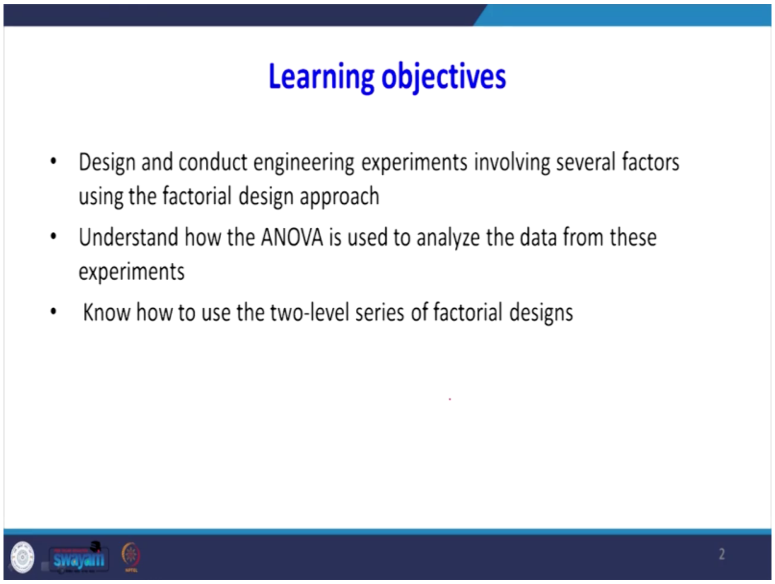
Data Analytics with Python
Prof. Ramesh Anbanandam
Department of Management Studies
Indian Institute of Technology – Roorkee

Lecture – 27
2 Way ANOVA

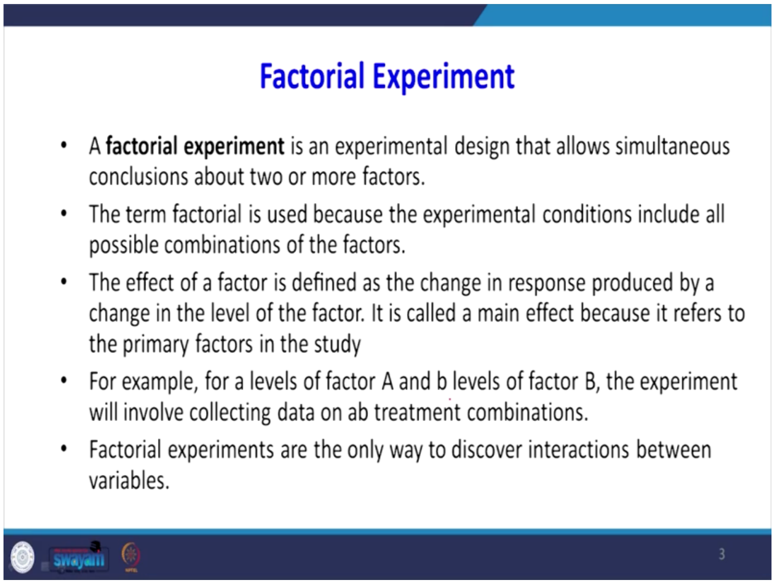
Dear students in the previous class we have seen randomized block design. In this class we will go to the next topic that is factorial experiments are 2 way ANOVA.

(Refer Slide Time: 00:37)



The learning objectives are designed and conduct engineering experiments involving several factors using factorial design approach, understand how the ANOVA is used to analyze the data from these experiments and know how to use 2-level series of factorial design.

(Refer Slide Time: 00:58)

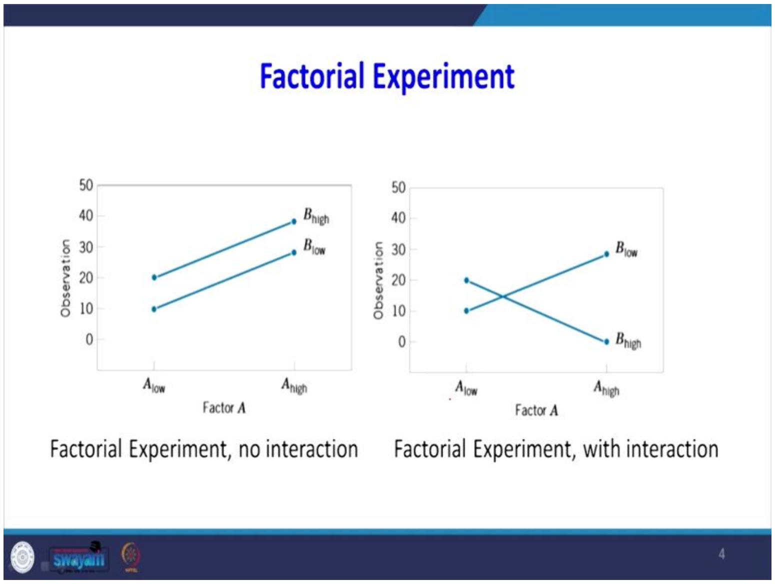


Let us see what is a factorial experiment a factorial experiment is an experimental design that allows simultaneous conclusions about 2 or more factors the previous 2 problems we did not see simultaneous effect of 2 variables at a time. The first problem when we whenever doing CRD completely randomized design we have taken only one independent variable the randomized to block design we have taken one independent variable and one blocking variable there was no interaction.

But in this lecture we will see that if there are 2 independent variable how there is a possibility of interactions we are going to see the effect of interaction also. The effect of factor is defined is the change in response produced by a change in the level of factor it is called main effect because it refers to the primary factors in the study. For example in levels of factor a and b levels of factor b small a is the level of factor a there may not be factor a there may be a levels small a levels may be low high there is a 2 level.

For factor b also there may be 2 level low medium high for example 3 level also possible that is small b the experiment will involve collecting data on a b treatment combinations small a and small b. Factorial experiments are the only way to discover interaction between variables.

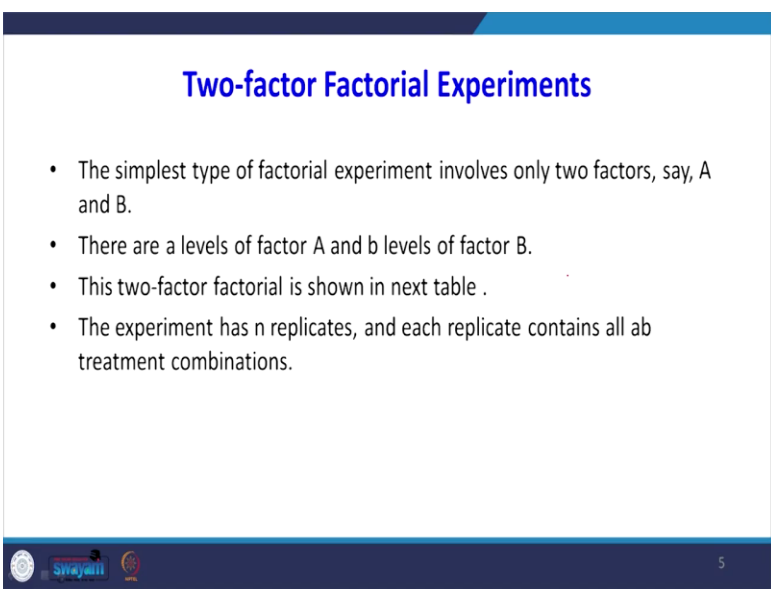
(Refer Slide Time: 02:27)



When you look at this diagram you see that there is a left side there is a factor that has 2 level low and high in observations you see the b and the factor they are also low a high, so these lines are parallel so these lines are parallel so that means that there is no interaction. When you look at the other side when the factor a goes low level to high level so what is happening there is a the crossing there is a interaction instead of a when a goes b is high when he is high then we also should be high.

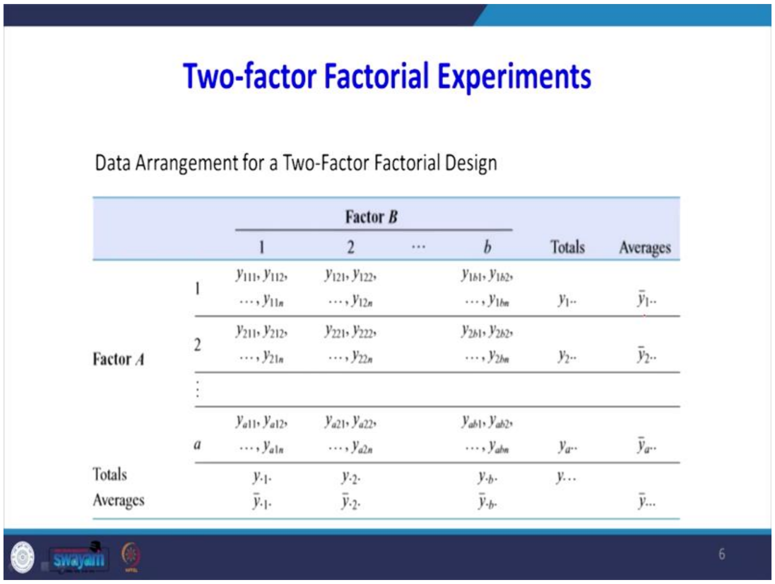
But now b has become it comes on the lower side so whenever there is a intersection that means that there is an interaction effect is there.

(Refer Slide Time: 03:15)



The simplest type of factorial experiment involves only 2 factors say A and B there are a levels small a, a levels of factor A and b levels of factor B this 2 factor lay shown in the next table the experiment has n replicates and each replicate contains all a b treatment combinations.

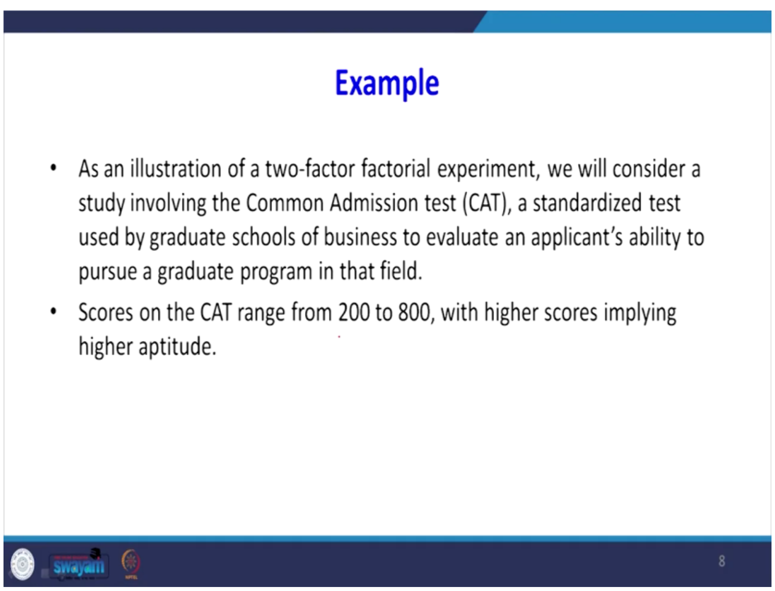
(Refer Slide Time: 03:38)



		Factor B 1	2	⋯	b	Totals	Averages
Factor A	1	$y_{111}, y_{112}, \ldots, y_{11n}$	$y_{121}, y_{122}, \ldots, y_{12n}$		$y_{1b1}, y_{1b2}, \ldots, y_{1bn}$	$y_{1\cdot\cdot}$	$\bar{y}_{1\cdot\cdot}$
	2	$y_{211}, y_{212}, \ldots, y_{21n}$	$y_{221}, y_{222}, \ldots, y_{22n}$		$y_{2b1}, y_{2b2}, \ldots, y_{2bn}$	$y_{2\cdot\cdot}$	$\bar{y}_{2\cdot\cdot}$
	⋮						
	a	$y_{a11}, y_{a12}, \ldots, y_{a1n}$	$y_{a21}, y_{a22}, \ldots, y_{a2n}$		$y_{ab1}, y_{ab2}, \ldots, y_{abn}$	$y_{a\cdot\cdot}$	$\bar{y}_{a\cdot\cdot}$
Totals		$y_{\cdot 1\cdot}$	$y_{\cdot 2\cdot}$		$y_{\cdot b\cdot}$	$y_{\cdot\cdot\cdot}$	
Averages		$\bar{y}_{\cdot 1\cdot}$	$\bar{y}_{\cdot 2\cdot}$		$\bar{y}_{\cdot b\cdot}$		$\bar{y}_{\cdot\cdot\cdot}$

Look at this there is a factor, factor here there is a, a level is there factor b there are b level is there so maybe observations will be there the observation in a jth cell for the kth replicate is denoted by y ijk in performing the experiment the a b and observations would be run in the random order. Thus like a single factor experiment the 2 factor factorial is a completely randomized design this is also kind of you CRD.

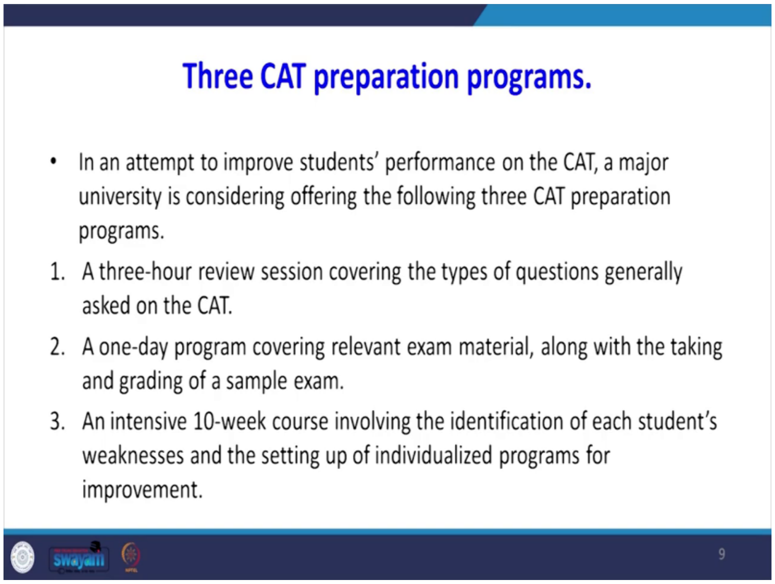
(Refer Slide Time: 04:10)



We will take an example with the help of example we will see how to do the 2 way ANOVA. As an illustration of your 2 factorial experiments we will consider a study involving Common Admission Test for example in MBA suppose we want to get admission MBA you have to go this Common Admission Test. A standardized test used by Graduate School of Business to evaluate the applicants ability to pursue a graduate program in the field.

Scores on the CAT range from 200 to 800 in India it is seen it is expressed in terms of percentile assume that the range is 200 to 800. So, that means the minimum qualify marks for CAT is a in terms of absolute term say 200 not in terms of percentile see but the higher scores imply higher aptitude.

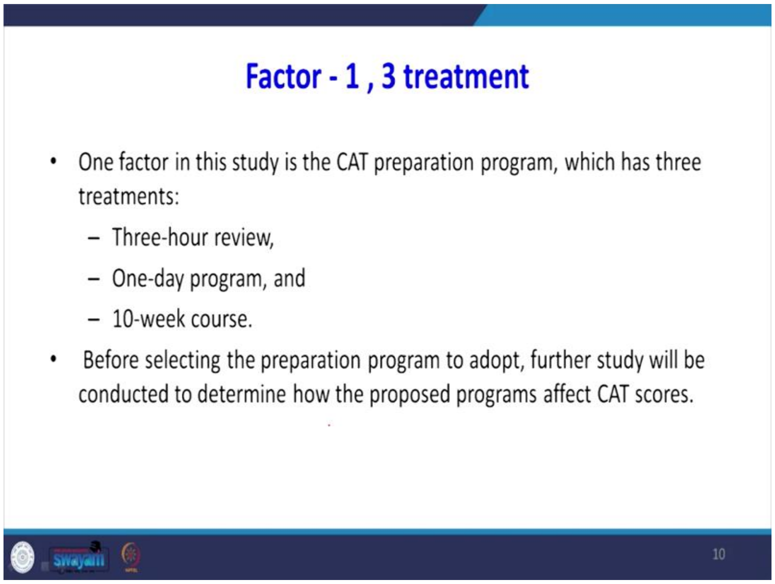
(Refer Slide Time: 05:04)



There are 3 CAT preparation programs in an attempt to improve students performance on the CAT a major university is considering offering the following 3 CAT preparation programs, there are 3 CAT population program the first program is 3 hour review, the second is one-day program the third one is intense you 10 weeks course involving. There are 3 type of coaching technique one is a 3-hour review session covering the types question generally asked in the CAT.

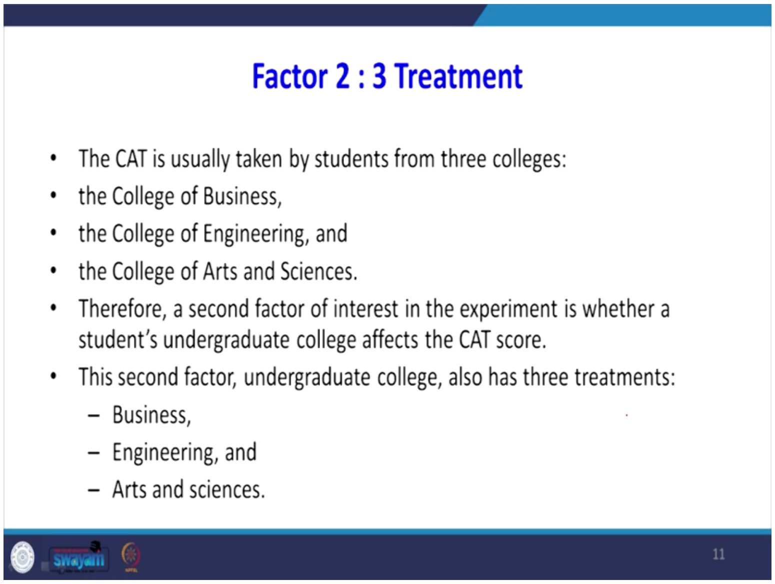
One day program covering relevant exam material along with; the taking and grading of sample exam. And intensive 10-week course involving the identification of each student's weaknesses and setting of individualized programs for improvement.

(Refer Slide Time: 05:04)



One factor is in this study is the CAT preparation program which has 3 treatment we are calling it as a treatment 3 hours review, one day program, 10 week course. There are 3 treatment before selecting the preparation programs to adopt further study will be conducted to determine how this proposed program effect to the CAT score. So, we are going to see there are 3 way of learning are preparing for the CAT examinations we are going to see the effect of these learning methods and how it is going to affect the performance in the CAT examination that is a CAT scores.

(Refer Slide Time: 06:34)

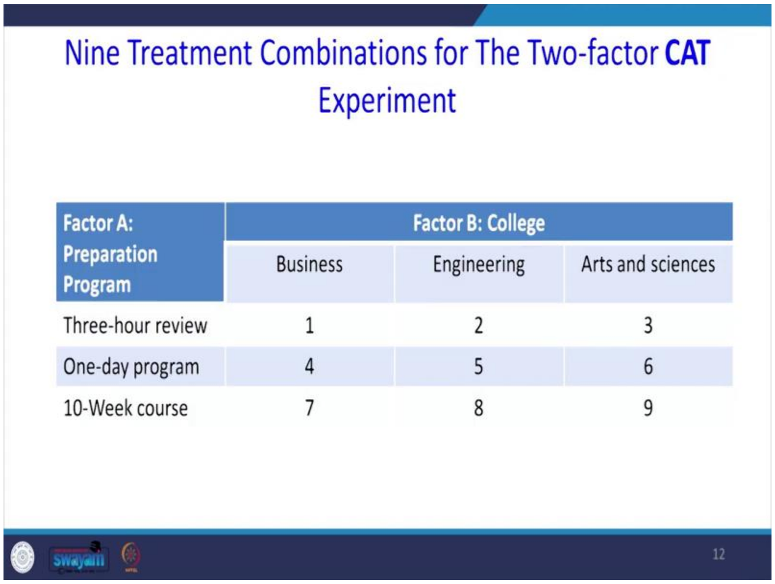


Factor 2 also there are 3 treatment CAT is usually taken by students from 3 colleges the college those have undergraduate business school, the college who come from engineering backgrounds, the College of Arts and Sciences. Therefore a second factor of interest in the experiment is whether students undergraduate college effect to the CAT score. Now for example in many IITs the arts and science students are not allowed to take MBA examinations but I would prefer at present many IITs they are allowing even our students also to get admitted into MBA program but they can take the CAT examinations.

Therefore we look at the will continue this problem therefore we a second factor of interest in the experiment is whether the students undergraduate college affect to the CAT score. The second factor undergraduate college also has a 3 treatment business a student may have Business Studies background or engineering background or art and science background. So, what we are going to study we are going to learn from this whether their undergraduate college or their background will affect their performance in the CAT's score.

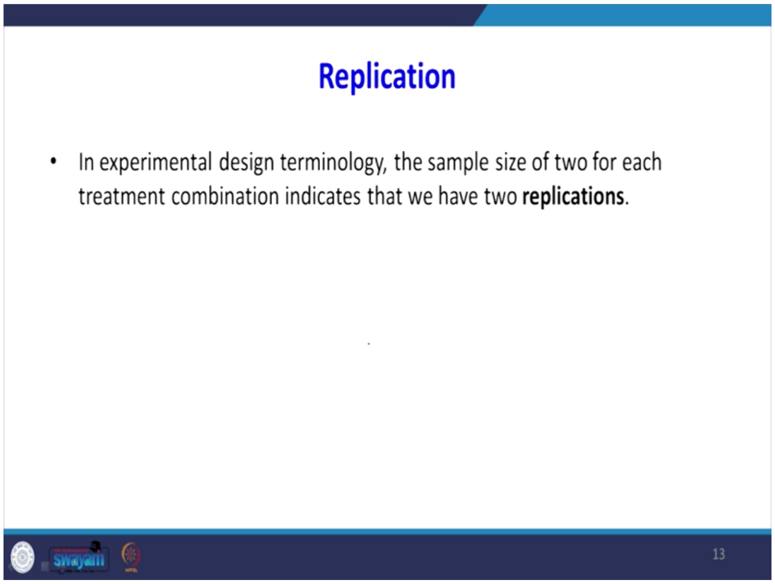
Maybe sometimes the engineering students do better in CAT examinations, sometimes the background from the BBA students in the business background may do the CAT examination better, sometimes the arts and science students there may be a possibility because they may not come across many quantitative subjects in their undergraduate they there is a perception that they may not do well in the examination that we will see in this exam in this problem.

(Refer Slide Time: 08:21)



So, what was done there are it is written in the table format in row it is taken the preparation program in factor in the column we take in college. So, what this table represents 9 treatment combinations for 2 factor CAT experiment, one factor is preparation program another factor is college there belongs to. So, you see that a person may be business background he may take 3 hours review he may be engineering background he may take 3 hours review he may be art and science background 3. So, there are 9 combinations are possible.

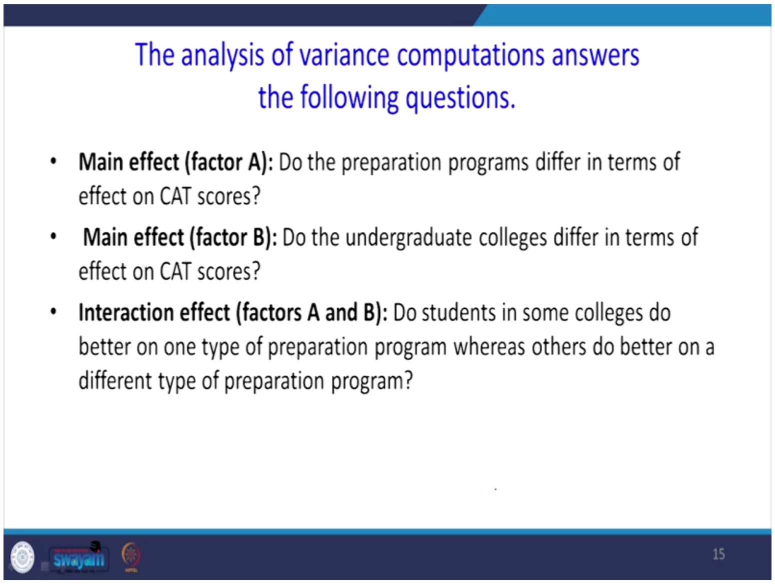
(Refer Slide Time: 09:04)



What is the replication an experimental design terminology the sample size for 2 for each treatment combination indicates that we have 2 replicates, if there are saying 2 for example here this 580 is the replication there are 2 students. So, what do you done a person who belongs to business background when he undertake 3 hours review of taking coaching method, so what was their marks. So, we have subject for 2 students we are subjected to 2 students.

Similarly 2 students are taking those who are engineering background and 3 our reviews, so how many 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 there are 18 observations.

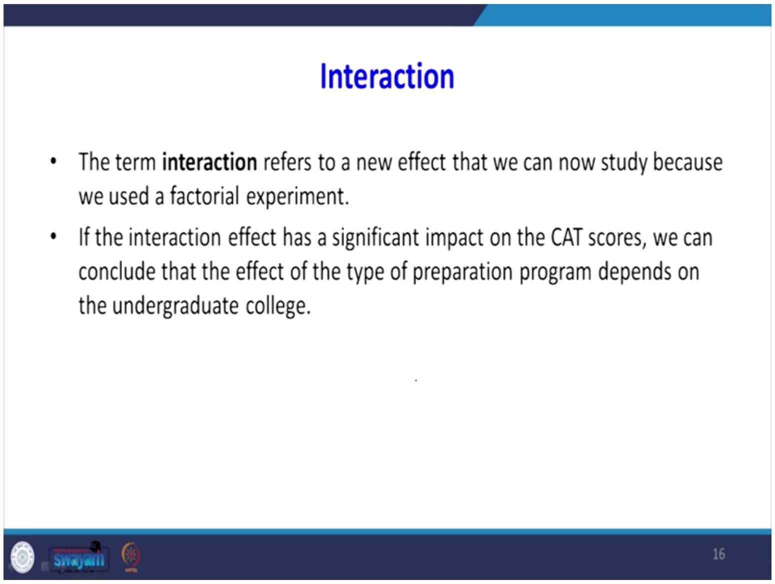
(Refer Slide Time: 09:55)



So, what is the analysis of variance computations answers the following questions what we are going to do that was the very important this lecture. So, what is the main effect factor a do the preparation programs differs in terms of effect on CAT's score. So, what we are going to see whether the different preparation programs affect their performance in the CAT course. Main effect B there is a factor B do the undergraduate college differs in terms of effect on CAT's course whether their undergraduate background is going to affect their performance in CAT's course are not.

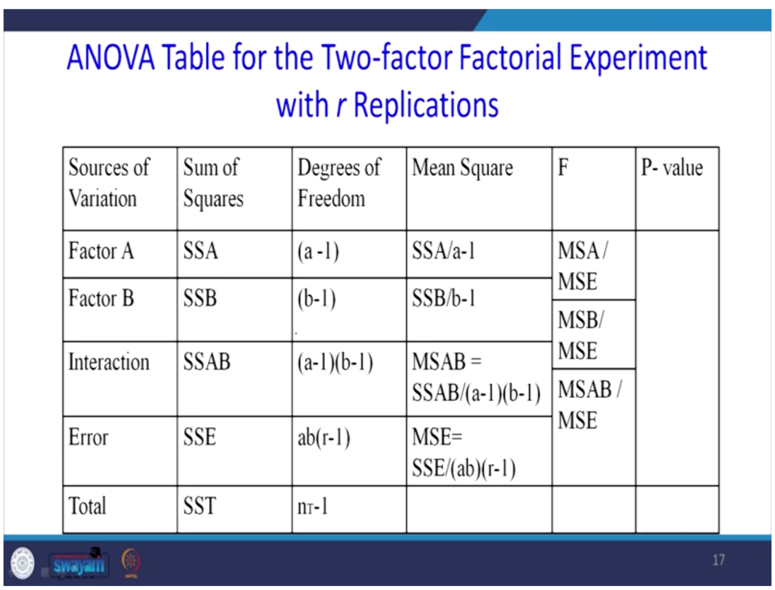
Then interaction effect that is a factor A and factors B do students in some colleges do better than one type of preparation program whereas others do better on a different type of preparation programs that we are going to see interactions.

(Refer Slide Time: 10:54)



The term interaction refers to a new effect that we can now study because we used a factorial experiment if the interaction effect has significant impact on the CAT's course we can conclude that the effect of the type of preparation programs depends on the undergraduate college that is the learning. If the interaction is significant we can conclude that the type of preparation programs depending upon the their undergraduate college background.

(Refer Slide Time: 11:29)



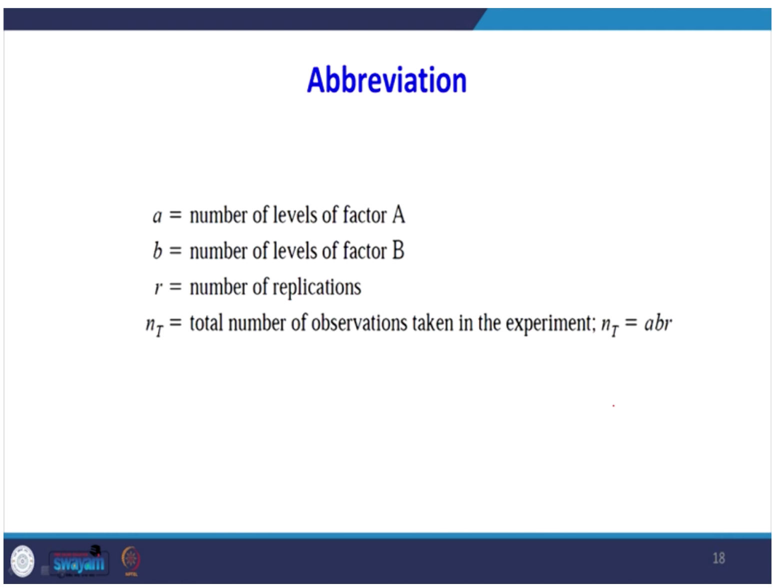
Sources of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F	P- value
Factor A	SSA	(a -1)	SSA/a-1	MSA / MSE	
Factor B	SSB	(b-1)	SSB/b-1	MSB/ MSE	
Interaction	SSAB	(a-1)(b-1)	MSAB = SSAB/(a-1)(b-1)	MSAB / MSE	
Error	SSE	ab(r-1)	MSE= SSE/(ab)(r-1)		
Total	SST	n_T-1			

This is a 2 factor factorial experiment setup so Factor A is here SSE a sum of square for factor A, SSB sum of square for factor B SSE is sum of square for error so when you subtract it this is AB for interactions. So, how to know the how to write the degrees of freedom so n T - 1 that is a degrees of freedom for total number of a levels in a a - 1 degrees of freedom for factor b - 1 for factor B when you multiply a - 1 into b - 1 degrees of freedom for interaction ok.

Then you see that this is a mean square for factor a mean square factor B this is mean square for interaction this is a mean square error. So, what we are going to see we are going to see effect of factor A and effect of factor B and effect of interactions. If you want to know the effect of factor a we have to write mean sum of square for factor A divided by MSE if you want to know the effect of B means sum of square for factor B divided by MSE.

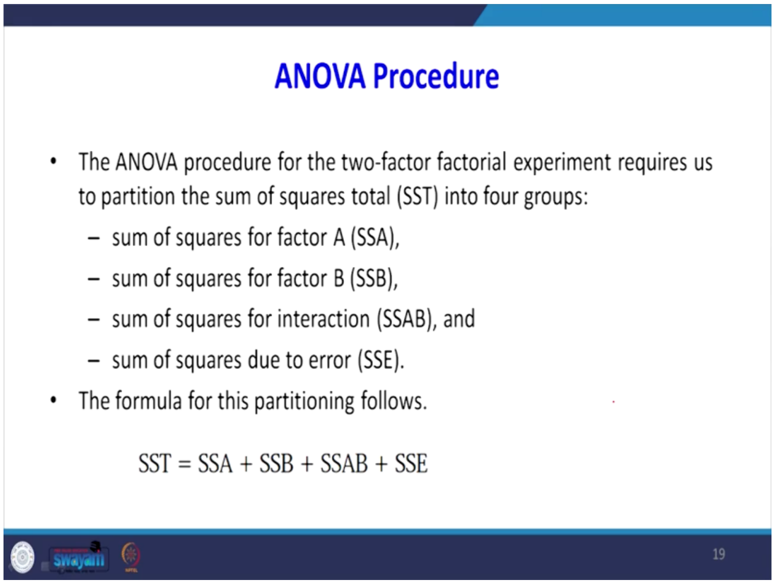
If we want to know the effect of interaction means sum of square MSA divided by MSE you see that in the denominator always there is a error term. Many students will do mistakes there because MSA divided by MSE the denominator always there should be a error term.

(Refer Slide Time: 12:49)



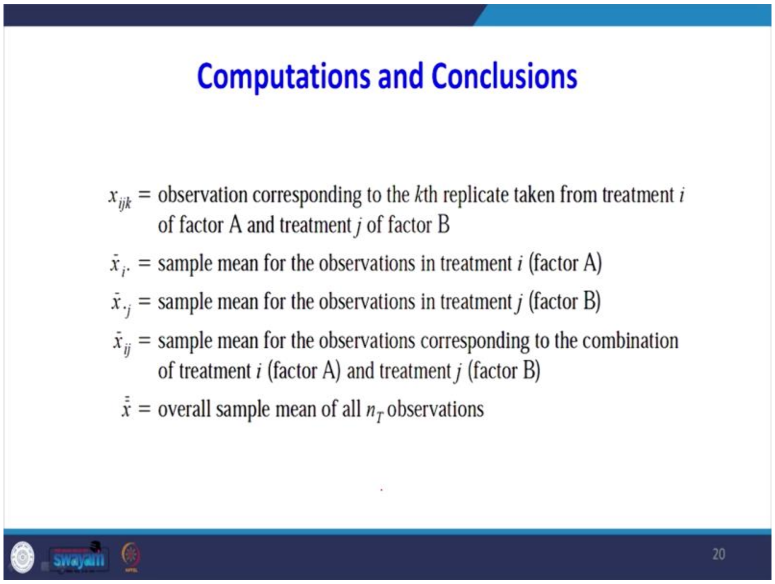
So, a represents number of levels of factor A, b represents number of levels of factor B, r represents number of replications nT represents total number of observations taken in the experiment. so, nT equal to abr because number of levels in a number of levels b material by number of applications.

(Refer Slide Time: 13:11)



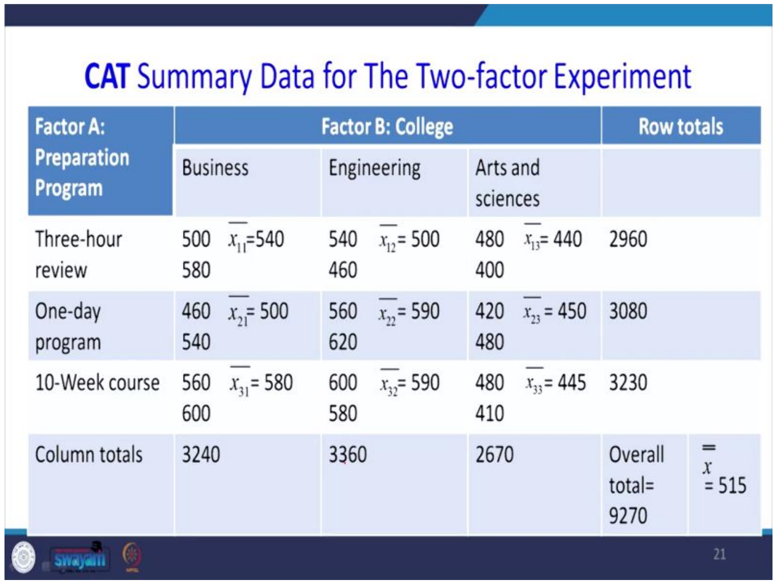
The ANOVA procedure for 2 factor factorial experiment requires us to partition of sum of squares into 4 groups, sum of square so SST we are partitioning into sum of square due to factor A sum of square for factor B sum of square for interaction of interaction and sum of squares due to the error term so, the formula for this partition follows SST equal to factor A sum of square plus factor B sum of square plus factor AB sum of square that is interaction sum of square plus error sum of square.

(Refer Slide Time: 13:48)



The notations are X ijk observations corresponding to the kth replicate taken from the treatment i of factor and treatment j of factor bXi dot bar represents sample mean for observation treatment i X dot j bar represents sample mean for observation in treatment j factor B, X ij bar represents sample mean for the observations corresponding to the combination of the treatment i factor a and treatment j factor B X double bar is overall sample mean of all nT observations.

(Refer Slide Time: 14:33)

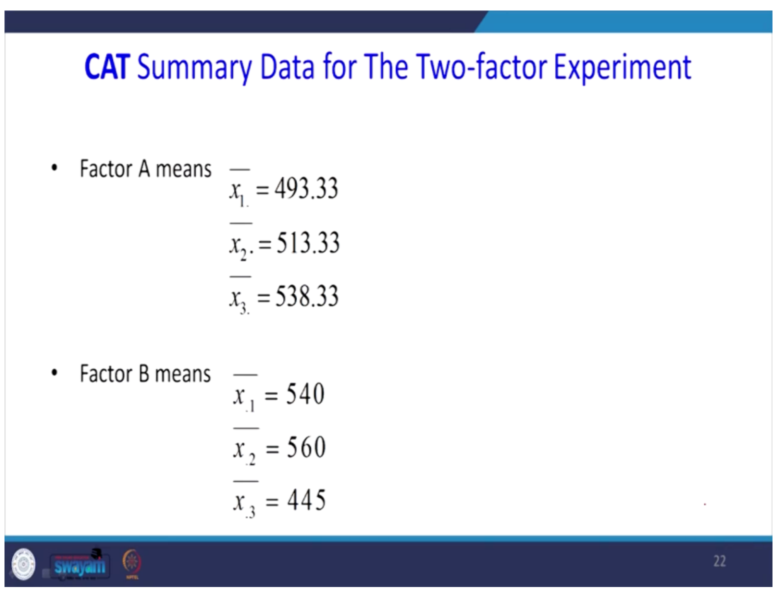


CAT Summary Data for The Two-factor Experiment

Factor A: Preparation Program	Factor B: College			Row totals	
	Business	Engineering	Arts and sciences		
Three-hour review	500 580 $\bar{x}_{11}$=540	540 460 $\bar{x}_{12}$= 500	480 400 $\bar{x}_{13}$= 440	2960	
One-day program	460 540 $\bar{x}_{21}$= 500	560 620 $\bar{x}_{22}$= 590	420 480 $\bar{x}_{23}$ = 450	3080	
10-Week course	560 600 $\bar{x}_{31}$= 580	600 580 $\bar{x}_{32}$= 590	480 410 $\bar{x}_{33}$= 445	3230	
Column totals	3240	3360	2670	Overall total= 9270	$\bar{\bar{x}}$ = 515

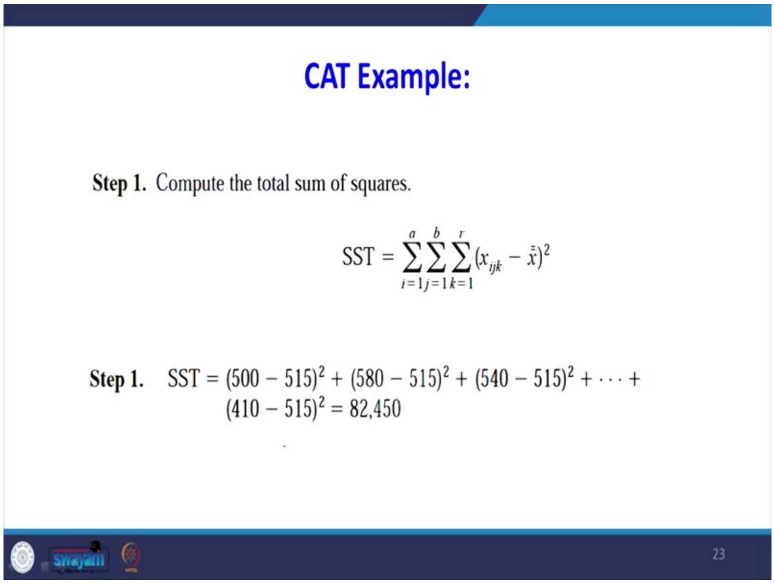
So, the first step is individually for each cell we have to find out the mean so here this look this location is X 1 1 bar mean is 540, X 1 2 bar mean is 500, X 1 3 mean is 500 the road total is 2960. So, the 2 1 mean is 500, 2 2 mean is 590, 2 3 mean is 550. So, third 1 X 3 1 bar equal to 580 X 3 2 590 X 33 445 the overall sum is 9270 the overall mean is 515.

(Refer Slide Time: 15:12)



Now the factor A means for row 1 X 1 bar equal to 493.33 X 2 dot bar is 513.33 X 3 dot bar equal to 538.33 for factor B means X you see that if it is a B only X dot 1 bar is 540 X dot 2 bar is 560 X dot 3 is 445.

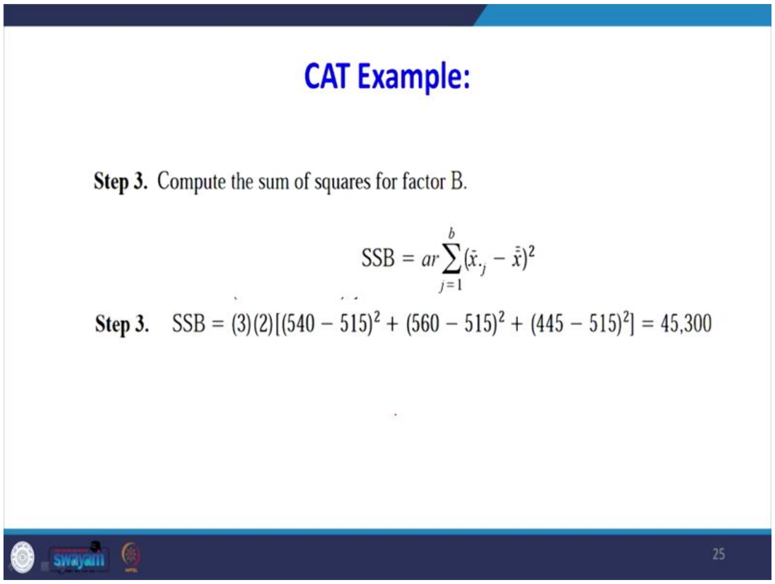
(Refer Slide Time: 15:45)



First step is to find out total sum of square we know that total sum of square is X ij k each element minus the overall mean whole square. So, what will happen we have to do for all the observations so that is coming 82,450. Step 2 compute the sum of square for factor A, so you write if you are writing you look at this if you are writing SSA so b into r summation of X i dot bar that is we can say it is a row mean row mean minus overall mean whole square.

So SSA is 3 b equal to 3 replicas 2 because 2 data set so 493.33 is row 1 mean minus overall mean whole square plus 513.33 is row 2 mean 515 whole square plus 538.33 minus overall mean whole square that is 60100.

(Refer Slide Time: 16:56)

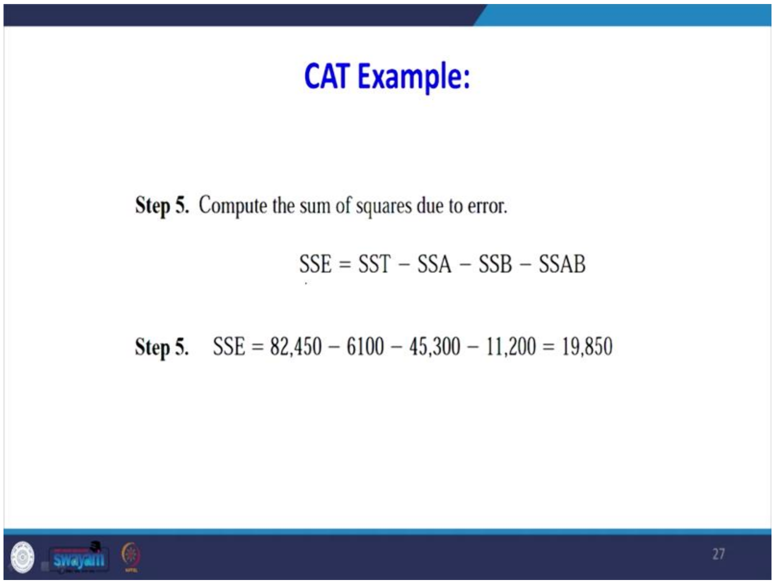


Now compute the sum of square of factor B whenever you write SSB see that a will come here ar summation j equal to 1 to b is nothing but your column mean, so column mean is 540 I am going back how we got the 540 and going back so 540, 580 this 580 540 560 445. Now let us go back next we are going to compute the sum of square for factor B so SSB equal to a our summation j equal to 1 to be X dot J bar - X double bar whole square.

So there are a is 3 to 2 replications 540 is your column mean and going back how we got 540 560 445 and go back this is 540 560 445 that was got this one that is why we got this one, 540 560 so SSB is 45300. Next we will go for SSAB compute the sum of square of interaction see here are in 2 there are 2 summations i equal to 1 to a j equal to 1 to b X ij bar minus that means in the cell what was the mean minus row mean minus column mean plus overall mean whole square.

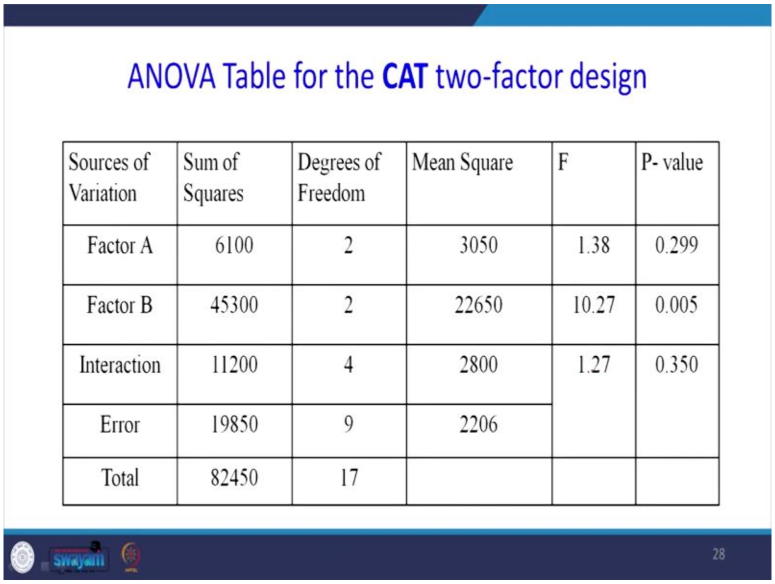
So, X aj is in that cell mean is 540 row mean is 493.33 minus column mean is 540 plus overall mean is 515 whole square. So, SSAB when you continue this we are getting 11200.

(Refer Slide Time: 18:48)



Then we will find out SSE, SSE is total if you subtract from the SST so total sum of square minus sum of square due to factor minus sum of square due to factor B minus sum of square due to a B so that SSE is 19,850.

(Refer Slide Time: 19:05)



Sources of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F	P- value
Factor A	6100	2	3050	1.38	0.299
Factor B	45300	2	22650	10.27	0.005
Interaction	11200	4	2800	1.27	0.350
Error	19850	9	2206		
Total	82450	17			

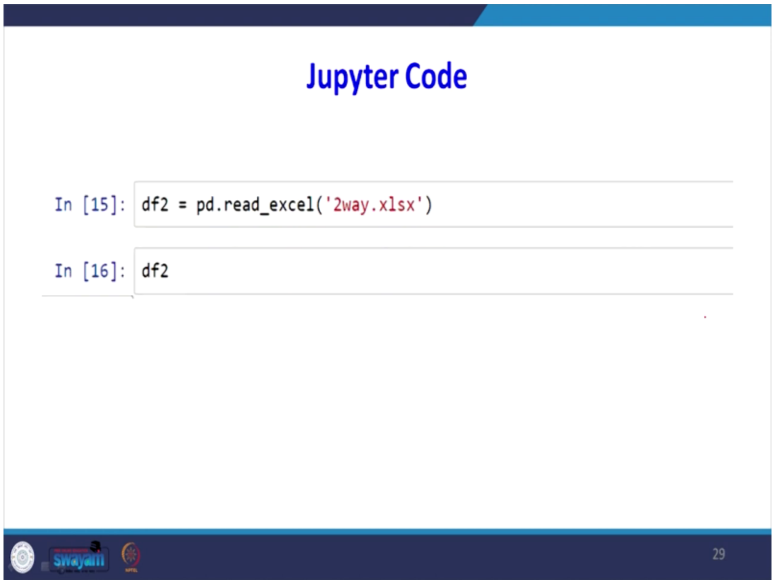
So, I have filled that value for factor a sum of 6100, 45300 interaction 11200 error 19850 so total sum of square is 82450, nothing but this 82450, 2450 is splitted into 4 parts one is due to factor A due to interaction in factor B and in due to interaction due to error. So, the degrees of freedom is there are 18 data set is there, so 18 - 1 is 17 there are 3 factors so 3 - 1 is 2 there are for factor A 4, for factor B there are 3 treatments so 3 – 2, 2 so we are getting a interactions it is number of levels in factor A that - 1 multiplied by number of levels in factor B that – 1.

So how we got this one go back you see that how we are getting the degrees of freedom for degrees of freedom for interaction right a – 1, a is number of level in factor a number of levels in factor that - 1 multiplied by number of levels in factor B that – 1. So, that value is this one you get 2 multiplied by 2 is 4. So, how we got this mean square 3050 when you divide 6100 by 2 3050 when you divide 43200 by 2 22650, 11200 divided by 4 for 2,800.

So 19850 divided by nine this one so how we got F value is when you divide 3052 by 2206 you will get to 1.38 when you divide 22000 divided by 650 divided by 2260 get 10.27 when you divide 2 2800 by 2207 is at 1.27 this is a corresponding p-value. So, what does happy here, here we are accepting null hypothesis when you accept a null hypothesis there is no effect of factor A.

Here interaction we are accepting null hypothesis there is no effect of interaction but there is an effect of factor B because it is less than point 0.05.

(Refer Slide Time: 21:46)

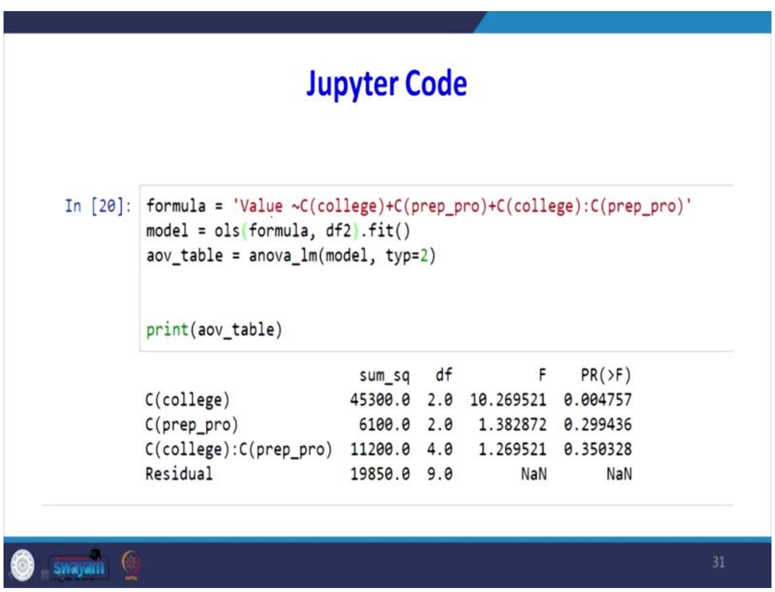


The data whichever is given there I have entered into the in Excel in excel file, so I am reading df 2 equal to pd dot read underscore Excel.

(Refer Slide Time: 22:02)

So, when I say df to see the data is in this permit value is in the first column there is 18 including 0 there are 18 values see the preparation program 3 hours up to this there are 3 hours, first is 6 data set this is one day this is 10 week, this is those who are belongs to business background, those belongs to engineering background, those who belongs to art and science background. Again this is for whenever one day preparation program who belongs to business background engineering background art science background when they go for 10 weeks intensive training program there also business background to engineering background and art and science background.

(Refer Slide Time: 22:55)



Here for formula equal to value tilde tilde College plus c preparation program plus c college colon c preparation program. This represents for interaction so model equal to ols that formula you can write it directly otherwise you can specify separately ols formula - df 2 - sorry dot fit so analysis of variance underscore table equal to anova underscore lm model c type minus when you write type 2 it is for 2-way anova.

Seriously when were remember for one-way anova here we write in type one we got one way ANOVA, so when you print on our table so we are getting this one. So, what is the meaning though the what we do it manually what you do in Python is same here what is happening this we accept a null hypothesis this we accept null hypothesis this we reject a null hypothesis. So, there is no interaction there is no effect of preparation program but there is an effect of the college background they belongs to.

They may be belongs to be engineering background they may be belongs to business background or art and science background because what we are concluding from here is that if there belongs to particular background there is courses their performance in CAT's score is otherwise we can say the college backgrounds affect their performance in the CAT's score it may be those who are belongs to engineering they can perform better or those who belongs to Arts and Science they may not perform better.

So what we are concluding here is the college they belongs to is an important variable on their performance in the CAT examination. We got the ANOVA table 2 way ANOVA table when we look at that see the preparation program is not a significant variable the interaction between college and the preparation program also not significant factor here but only the college there belongs to is an significant factor that means there are 3 possibility their college background may be one is Arts and Science second one is a business third one is engineering background. So that factor will affect their performance in the CAT's score.

Dear students in this class we have studied what is a 2 way ANOVA then we have taken one problem we have traditionally we have solved that 2 way ANOVA. Then I have explained the theoretical background behind this 2 way ANOVA then the same problem we have solved with the help of Python. Then we have interpreted the result. The next class will go to another topic that is a regression analysis because this analysis of variance and regression analysis are it is like a 2 side of the same coin.

Even ANOVA can be solved with the help of regression analysis even here aggression problem can be solved with the help of ANOVA. So, the next class I will meet you with another new topic called regression techniques, thank you very much.