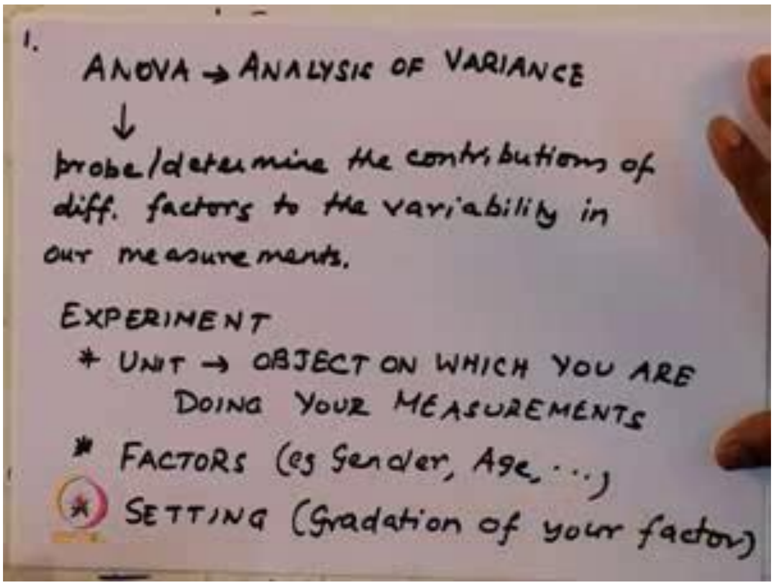
Introduction to Biostatistics
Prof. Shamik Sen
Department of Bioscience and Bioengineering
Indian Institute of Technology, Bombay

Lecture - 38
ANOVA

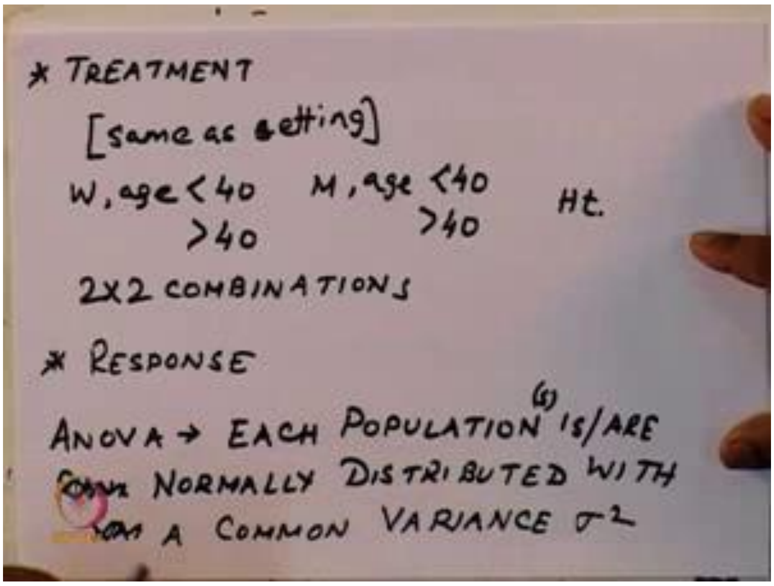
Hello and welcome to our lecture today. So, we would briefly recap what we had discussed in last lecture which was analysis of variance right or ANOVA.

(Refer Slide Time: 00:27)



So, ANOVA allows us to probe or determine the contributions of different factors to the variability in our measurements. So, when you do an experiment you have the experimental unit which is the object on which you are doing your measurements. You have your factors or example gender is a factor, or let us say height age can be a factor so and so forth. You have settings, which is the gradation of your factor. So, also something called treatment.

(Refer Slide Time: 02:40)



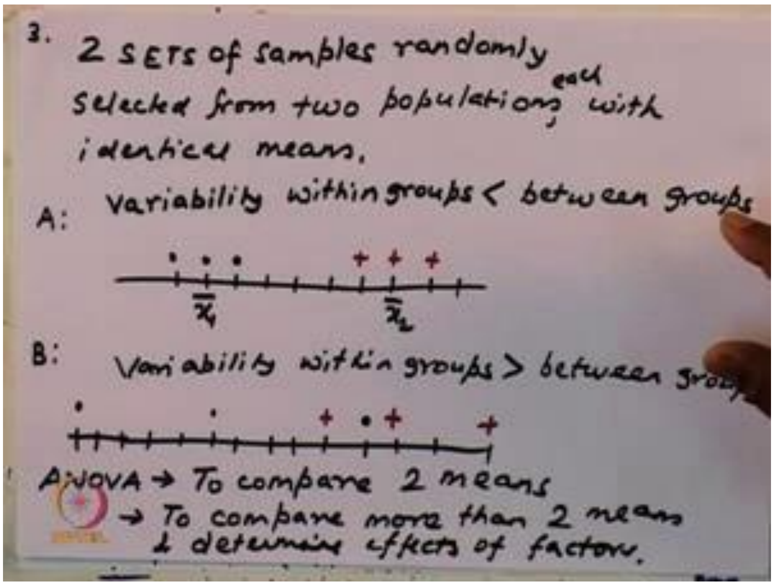
So, the treatment can be sometimes it can be the same as setting. So, let us say your setting is just men versus women or gender.

So, in case in that case your treatment is same as you are setting; however, let us say your treatment can be women with age less than 40 and greater than 40, similarly men with age less than 40 and greater than 40. So, you can add this. So, in this case you have 2 into 2 combinations or treatment settings. So, if let us say on top of that I introduce height as another metric then I have 2 into 2 into 2, 8 combinations. So, treatment may or may not be equal to the setting value. And finally, is your response is what you measure.

So, ANOVA allows you to try to understand the role or effect of each of these individual factors on the overall variability in the population. So, ANOVA is based on one assumption, that each population is normally distributed, with the common variance. See if there were more than population. So, it is each population if you have populations are normally distributed with the common variance sigma square. So, what you are saying is each population is normally distributed and the variance is the same for each population.

So, the means may or may not be equal to each other. So, let me give an example of how that might be.

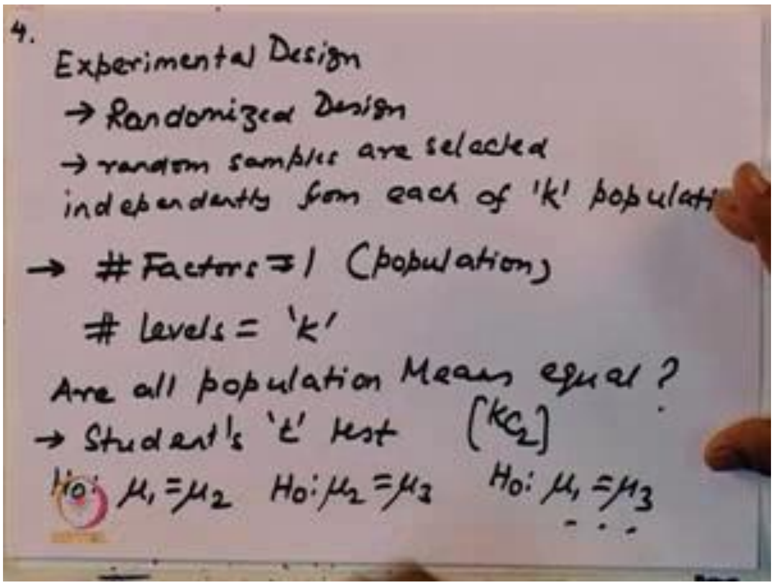
(Refer Slide Time: 05:26)



So, imagine you have 2 sets of samples randomly selected from 2 populations, with identical comma each with identical means. So, let me draw one particular case. So, imagine one populations; one population one of the samples gives you these value. The other of the samples other of the samples gives you these value. So, this is your scenario A.

Imagine the other scenario. So, this is your x 1 bar and this is your x 2 bar. The other case let us say you have one population here. So, these are your 2 situations. So, what you observe is in this situation A there is the variability within the groups is much less than between groups, versus in this case the variability within groups is much greater than that between groups. So, this is where the ANOVA approach is important because ANOVA can be used to compare 2 means. It can also be used to compare more than 2 means and determine effects of v s factors. So, in ANOVA you can do one of the 2. So, as we said let us take an example. So, there are various ways in which you can draw the sample. So, one of the experimental ways.

(Refer Slide Time: 09:01)

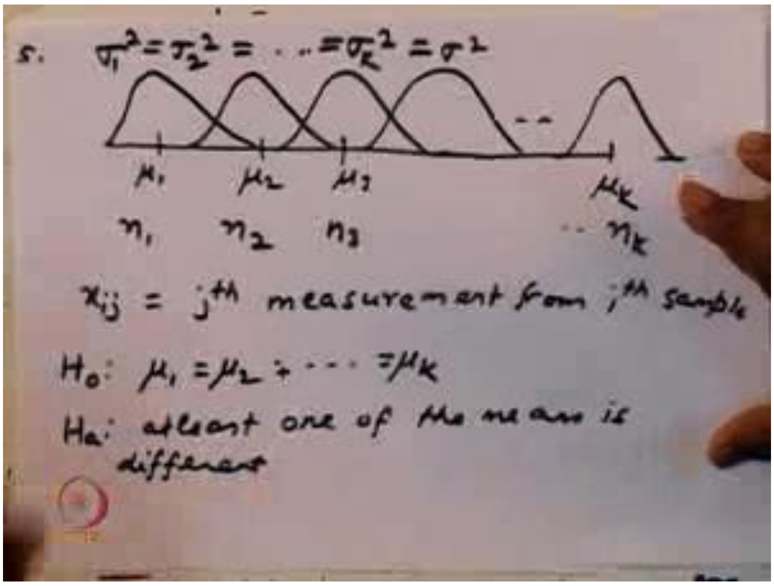


Designs is called the randomized design, design where random samples so are selected independently from each of k populations.

So, in this particular case the number of factors is equal to 1, since you just have population as your variable, but the levels. So, levels of the factor is equal to k because you have k such populations. So, you can ask the question are all population means same. So, it is possible to do the students t test. So, you can do the students t test, but then you will have to test various hypothesis, you have let us say mu naught mu 1 equal to mu 2 another high test of hypothesis, mu 2 equal to mu 3 H naught is mu 1 equal to mu 3 so and so forth.

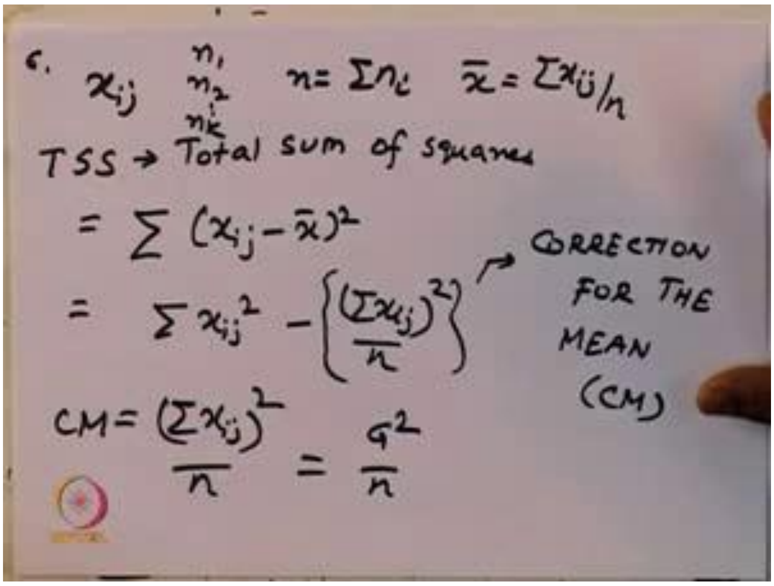
So, if you have k different populations you have k c 2 is a number of tests of hypothesis that you will have to perform. So, increase in the number of tests increases the possibility of error. So, ANOVA provides you that ability to come to this conclusion with a single test. So, what do we do in ANOVA?

(Refer Slide Time: 11:21)



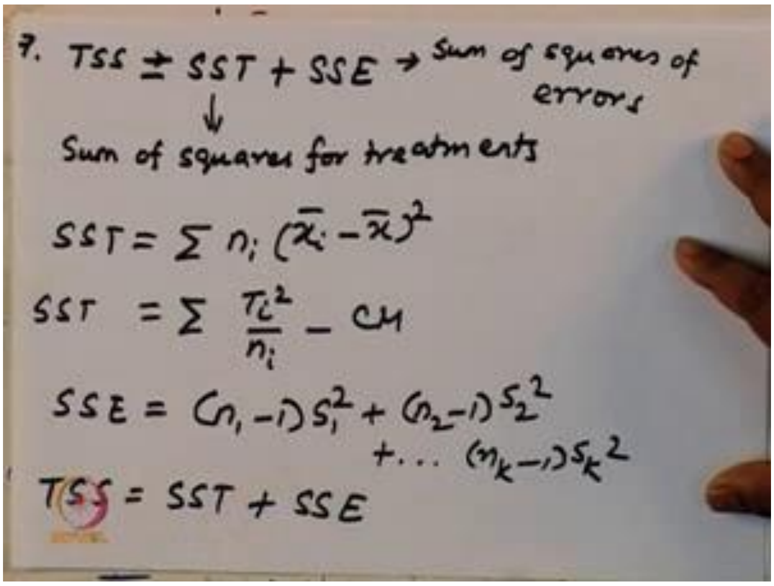
So, here you have mu 1, mu 2, mu 3, mu k each of these populations your variances are same. So, any want to test the heights. So, let us say your sample sizes are n 1, n 2, n 3, n k. And you can have xij is the jth measurement from ith sample. So, you want to test the hypothesis, that H mu 1 equal to mu 2 equal to mu 3 so and so forth. So, your null hypothesis is mu 1 equal to mu 2 equal to mu k, and alternate hypothesis is at least one of the means is different.

(Refer Slide Time: 13:09)



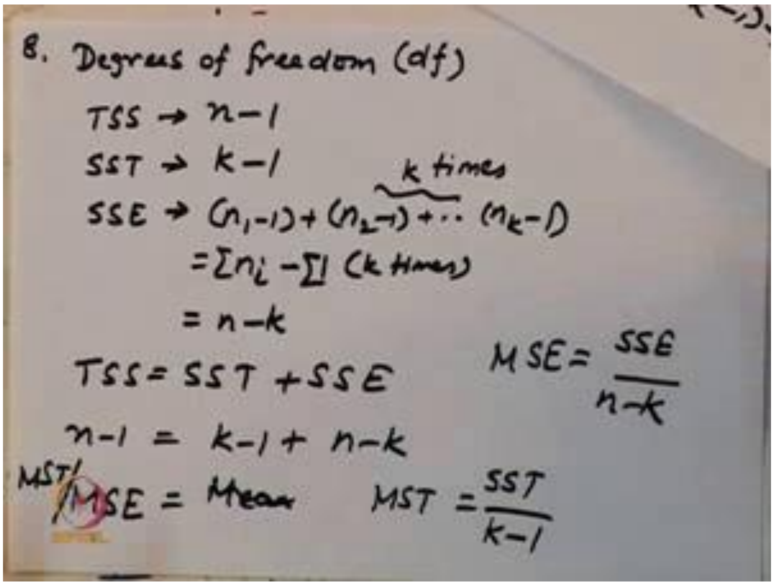
So, what you do? So, as I said xij is your jth measurement from ith sample. So, you calculate the following quantities TSS is total sum of squares. And you have samples drawn from n 1 n 2 dot dot n k you define n as summation of all n i. So, TSS is defined as total sum of squares, and let us say x bar is summation xij by n. So, TSS is defined as summation of xij minus x bar whole square. So, you can show if you expand this you can show this is nothing, but summation of xij square minus. So, this term is called the correction for the mean or CM. So, CM is summation xij whole square by n and this you can write as G square by n. So, G represents the basically sum of all the terms.

(Refer Slide Time: 14:44)



So, in ANOVA you distributed the TSS into 2 fragments, you distribute that TSS into 2 fragments one you call as SST. So, TSS equal to SST plus SSE sum of squares of treatments. So, this stands for sum of squares for treatments and this ter MST o sum of squares of errors. So, you define SST as summation of x i bar is the sample average for the ith sample n i is a sample size of the ith sample. So, this you can again expand and show you can show that SST is same as t i square by n i, t i is the sum of all in ith sample total sum of ith sample. So, an SSE is given by, so, you can show that TSS is equal to SST plus SSE. So, what can I write over the degree of freedom or degrees of freedom for each of these terms.

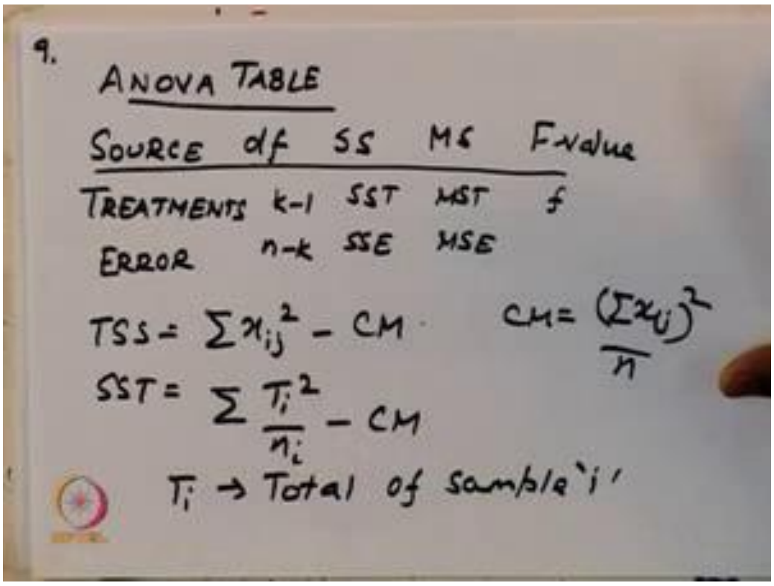
(Refer Slide Time: 16:36)



For TSS you have n terms which you square and add up. So, TSS has to be n minus 1 for SST you have k terms. So, this for SST you do this over all k terms. So, SST is the degree of freedom for SST is k minus 1. And for the errors SSE so, you haven 1 minus s 1 square plus so and so forth. So, your SSE has contribution forn 1 minus 1 plus n 2 minus 1 plus dot dot dot n k minus 1. So, this is k times. So, this you can simplify equal to n 1, So, summation n i you have n terms all together and minus 1 which is k times summation one k times this is nothing, but n minus k.

So, given that TSS is equal to SST plus SSE. So, degrees of freedom of n minus 1 is equal to k minus 1 plus n minus k. So, the degrees of freedom also add up. And for corresponding to these terms you can define MSE which is the mean square either you can define MST mean square for treatments or mean square for error it is. So, MST is defined as SST by this degree of freedom that is SST by k minus 1. And MSE is the mean square of error is given by SSE by n minus k.

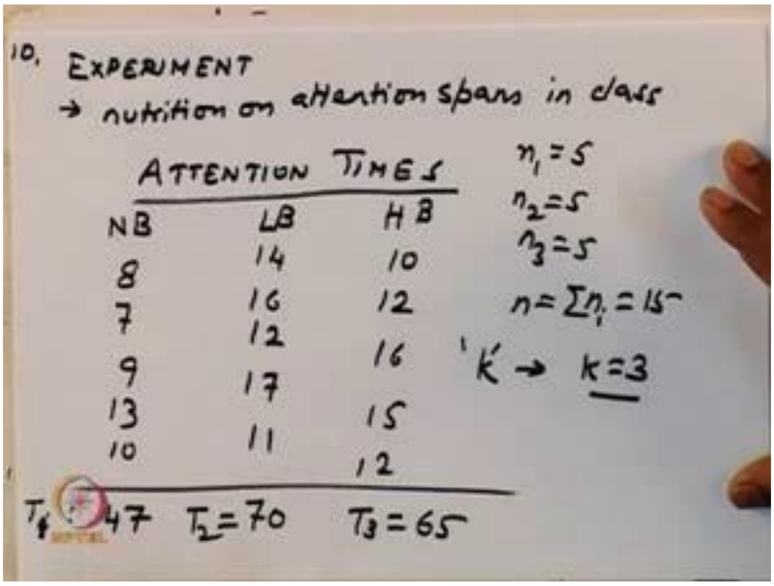
(Refer Slide Time: 18:50)



So, when you plot all these together you generate an ANOVA table which look something like this. Your typical ANOVA table will look something like this - source, degrees of freedom, sum of squares, mean squares and f value. Your source correspondence to each of the treatments, if you have 2 treatments you will have 2 if you have 4, you will have 4 of these conditions and you have accumulation from error is degrees of freedom is k minus 1 and n minus k. This is your SST or SSE your MST or MSE and you will get some value some f value.

So, again, you have TSS is summation xij square minus CM SST. So, CM is CM is given by summation xij whole square by n SST is summation of t i square by n i minus CM, where t i is a total of sample i. So, let us do a case.

(Refer Slide Time: 20:34)

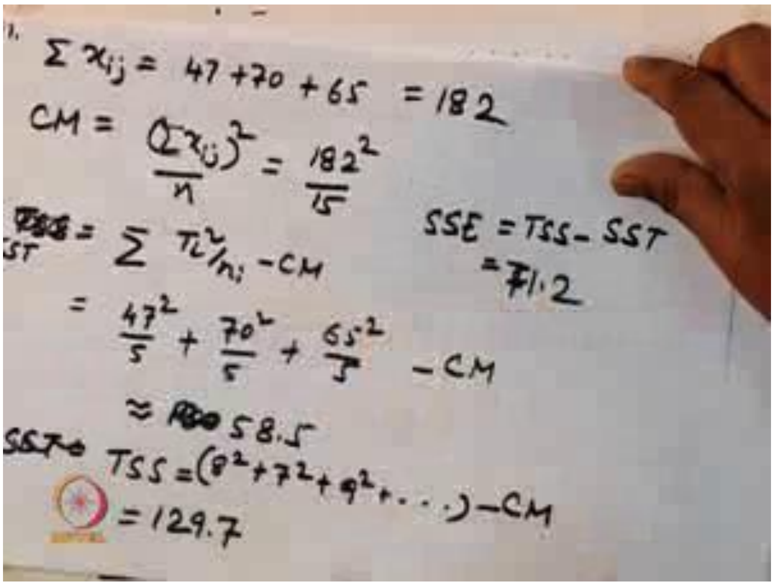


So, imagine, you are doing an experiment where you wish to study the effect of nutrition on attention spans in class. In other words, you want to see that is there a difference between students who have their breakfast and then come to class whether they pay more attention compared to students who have light breakfast or no breakfast. So, I have a plot of attention times and I have 3 categories. Let us say 3 treatments students who did not have any breakfast. So, no breakfast the values are who had light breakfast and who had heavy breakfast.

So, for each of these cases I can calculate t i. So, t i is what is total of sample i. So, you haven 1 it is a sample size for condition 1 which is equal to 5, n 2 is also 5 n 3 equal to 5. So, n is summation of n i is equal to 15. So, what is the value total sum of t i, you can sum these up. 15 this is 47, this is 70 and this, this is t 1 this is t 2 and t 3. I can find out as 65.

So, what is the k, k value is the number of populations. So, in our case k is equal to 3 you have 3 different populations. So, summation xij is you add up all these values. So, we should come out to be 47 plus 70 plus 65 equal to 182. So, your CM.

(Refer Slide Time: 23:01)

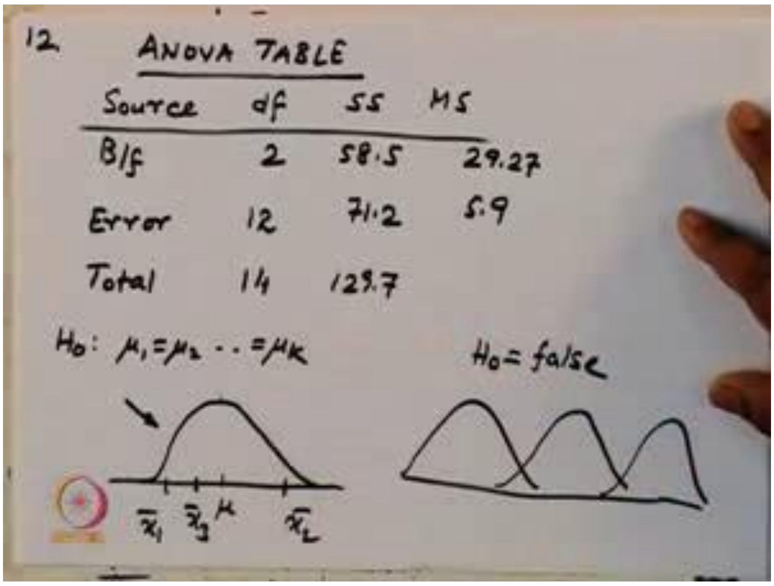


The correction for the mean is summation xij whole square by n should come out to be 182 square by 15. Your TSS is summation t i square by n i minus CM. So, which you can calculate as 47 square by 5 plus 70 square, 70 square by 5 plus 65 square by 5 minus, CM and you will get a value of this equal to roughly 130.

SST is given by summation t i square by n i minus CM, your TSS, So, this is actually gives you a value of SST will give you a value of roughly 55 58.5. TSS you have to add up all the squares.

So, in other words you have to do 8 square summation xij whole square. So, TSS is equal to 8 square plus 7 square plus 9 square you square up all the terms, in for each of these conditions and minus you do CM. So, this comes out to be value of 127.7. TSS returns your value of SST. So, you can calculate SSE is equal to TSS minus SST and this gives you a value of roughly 58.53, no SSE gives you a value of 71.2.

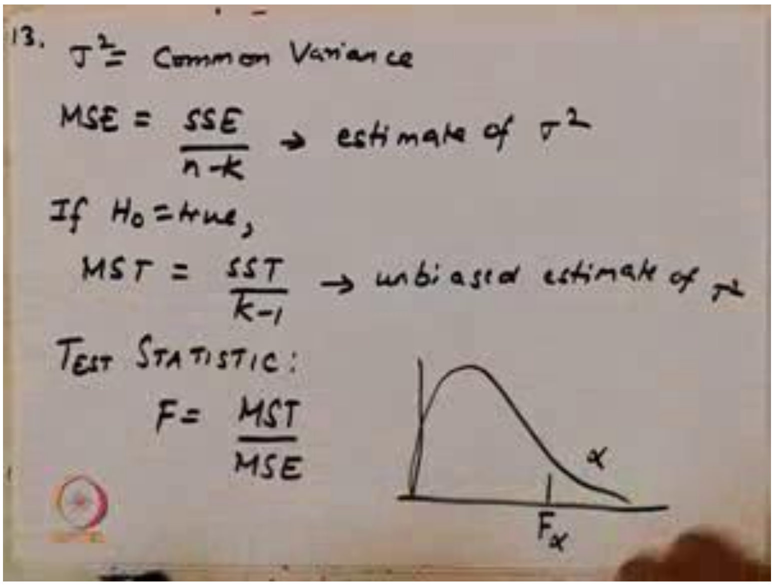
(Refer Slide Time: 25:13)



So, based on these values we can calculate the ANOVA table. So, your source is you know is your meal or breakfast your d f you have SS you have MS and you have error. So, for b f your degrees of freedom is 2 because k is equal to 3, for error you have equal to 12 n minus k will give you a value of n equal to 15. So, n minus k will give you a value of 12. This we calculate as 58.5 this you calculate as 71.2. You can accordingly calculate MS and ME and you can do the total this is 14 this is 129.7. So, this is your ANOVA table.

Now, what do you need to do to calculate your test your hypothesis? So, your H naught is mu 1 equal to mu 2 equal to mu k. So, if these means for all same then, you would have had a distribution let us say hypothetically, you can have x 1 bar here x 2 bar here x 3 bar here. So, in this case you might have agreed to say H 2 is true; however, if your case was something like this here your H naught is false.

(Refer Slide Time: 27:09)



So, sigma square your assumption is sigma square is common variance for all k. So, your MSE is given by SSE by n minus k. It is an estimate of sigma square and if your H naught was true. So, your MST, MST which is SST by k minus 1, it should give you an unbiased estimate of sigma square.

So, you can use the test statistic as f equal to MST by MSE. And as before you can use your f test calculate f of alpha. And then see whether you can test whether if your f value is greater than f of alpha then you know your hypothesis is not true.

With that I conclude my talk today. So, you get an idea of how ANOVA can be used instead of repeatedly using students to test for calculating whether means are same. You can have come to this conclusion using a single approach which is your ANOVA.

So, you create your ANOVA table by calculating by distributing the total sum of squares or TSS into SST or sum of squares of treatments and SSE which accounts for random error. So, sum of squares of errors from there you calculate the mean square error or the mean square sum of t and then use the statistic MST by MSE to find out whether your means are same or they are distinct.

Thank you for your attention.