Dealing with Materials Data: Collection, Analysis and Interpretation

Prof. Hina A Gokhale
Department of Metallurgical Engineering and Materials Science
Indian Institute of Technology, Bombay

Analysis of Variance - 1

Hello and welcome to the course on Dealing with Materials Data. We have come a long way, we have learned the basics of statistics, very, very basic of statistics, not in great details. Then we saw the one application area which is Regression Analysis and today we are going to have a session on Analysis of Variance.

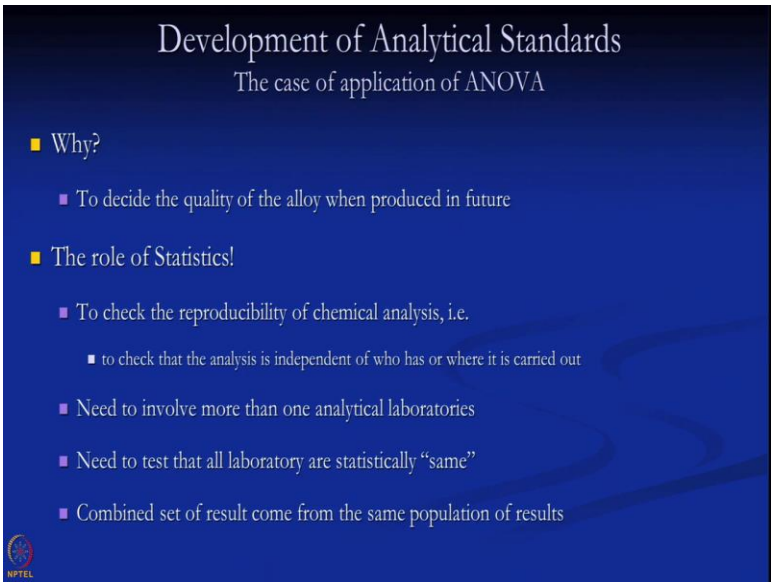
This time for a change we will start with a case study. We will actually take a problem which has been solved and we will see, how the analysis of variance was useful in doing that and what I would like you to appreciate is that whatever we go through, you will realize that it is a very natural process of doing it and it is not a sum process that has been devised to fit into the problem rather the problem itself gives a rise to this particular methodology and that is how analysis of variance we are going to introduce. Then we will develop it as a time goes in a more appropriate and theoretical manner. So, let us start.

I want to consider the case of development of analytical standards, when a new alloy is produced, it is very important for the producer, the industry to know, that the alloy that is produced has the same chemical values. Now what happens is that, a industry makes a alloy and it sells it to another company.

Now the industry has worked out its chemical standard in their own chemical laboratory and has told the customer that this is my chemical values of this alloy, this is the chemical composition of this alloy. Now the customer takes a sample and tests in his or her own laboratory and he finds it different. Then whom to consider correct?

In other words what I am trying to say, is that, when a chemical standard is decided, when a producer says that I am giving you an alloy with this chemical standard, it means that it should not have any effect of where it is done, who is doing it, who has made the samples and which testing machines have been used.

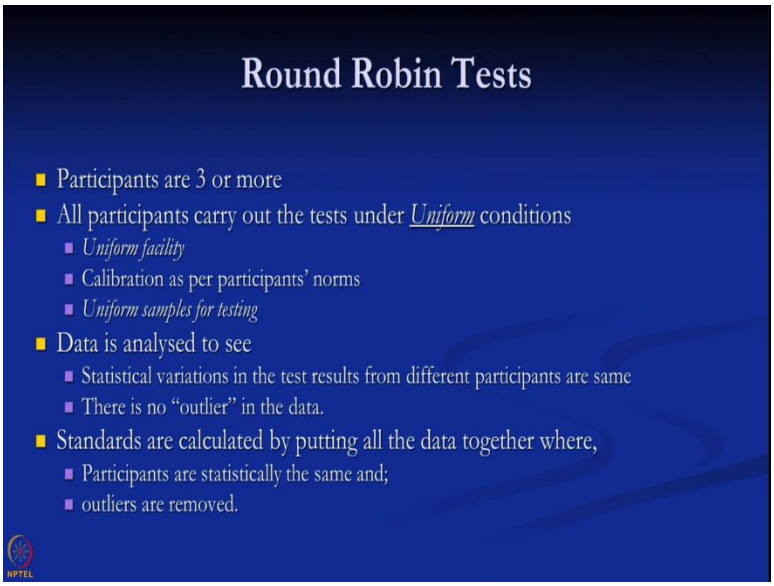
(Refer Slide Time: 03:28)



In order to do this, what we really want to see is the reproducibility of the chemical analysis, to check that the analysis is independent of who or where it is being done. If you want to do this, to begin with we must involve more than one chemical testing laboratories to test the samples and then we need to see, statistically test, that all samples are giving the same values.

If it is the case, if we find that we did the test in different chemical laboratories and all laboratories' results are statistically same. Then we can take the combined set of all the chemical standards, chemical analysis that has been done from all the laboratories and come up with a standard. This procedure is called Round Robin Tests in Statistics.

(Refer Slide Time: 04:32)

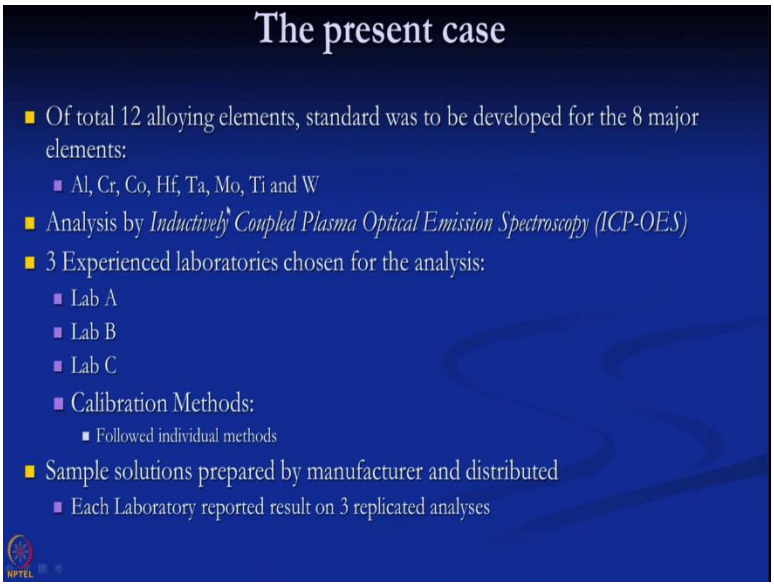


Where we expect that there are at least more, three or more participants in the Round Robin Tests. All participants carry out test under a Uniform condition. The facility is uniform, the calibration is as per the participant's norm and the uniform samples are given for testing.

Then the data is further analyzed to see that the statistical variation in test results are not due to the different laboratories, but they are due to the simple statistical variations. There are no extreme values in it and then we finally come to a standards by putting all the data together where all the participants are statistically same and then, in this case at least we remove the outliers.

Please remember our earlier discussion on outliers which are the extreme values. There we had said that blindly removing the outliers is not a solution. But in this particular case, the way has been to remove the outliers and then calculate the chemical standards. So, let us look at the present case.

(Refer Slide Time: 05:51)

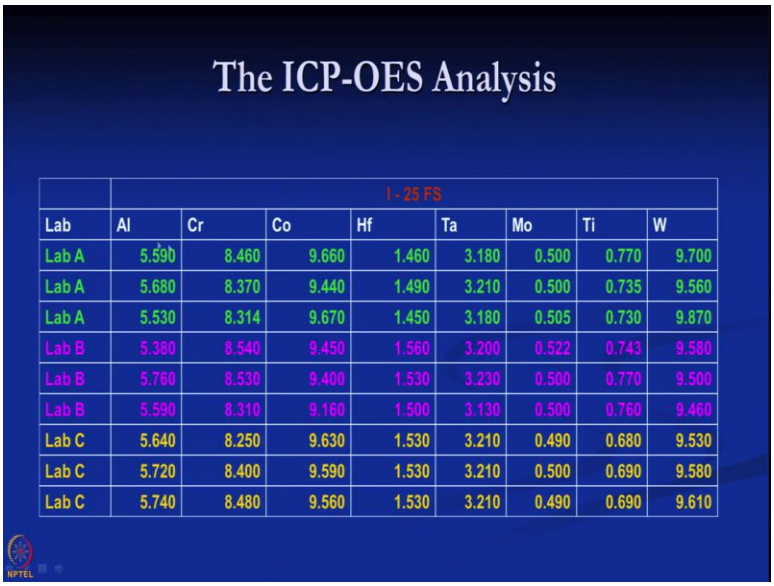


There is an alloy having a 12 alloying elements and we have to develop the standard for 8 major elements as given here as aluminum, chromium, cobalt, hafnium, tantalum, molybdenum, titanium and tungsten. The analysis it is decided that everyone should do analysis through a process which is called Inductively Coupled Plasma Optical Emission Spectroscopy, ICP-OES.

Three experienced laboratories were chosen for this analysis. Because you have to analyze 8 elements in a 12 alloying element alloy. So, it needs a considerable experience. So, three laboratories were chosen. Let us call them Lab A, B and C. The calibration method was left to the laboratory, they have to follow their own individual calibration standards and methods.

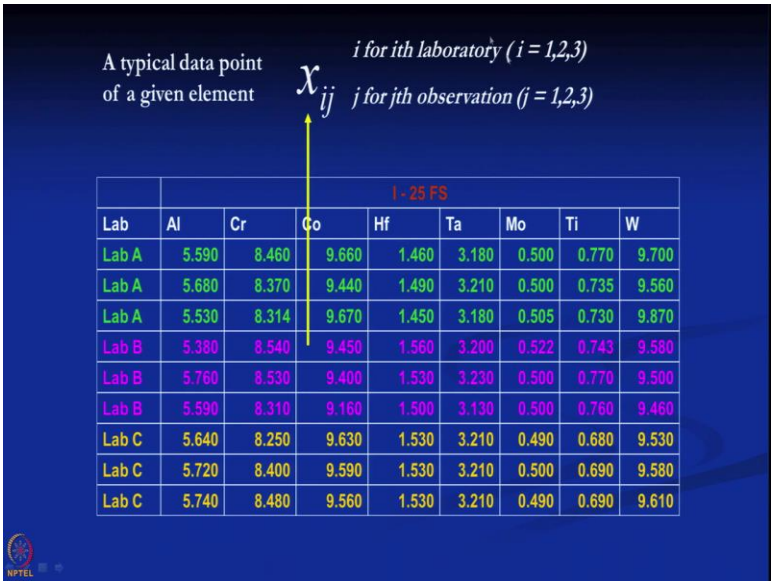
The sample solutions were prepared by the manufacturer and distributed to each laboratory and each laboratory had to report result on three replicated analysis. I am going to say more about this replica and repetition because there is a very common misunderstanding about the two. But we will talk about it when we talk about design of experiment. So, here I am saying that three replicated analysis, it means that not the three observations of the same analysis, you have to do three analysis itself that is called three replicated analysis.

(Refer Slide Time: 07:43)



See here I am giving you the table of the results for the 8 elements. This is in ppm. So, these are the results that laboratory A, which is in green, the other is in magenta and the third one is in yellow for ours easiness to understand it and these are the chemical analysis values they have found using the same ICP-OES analysis.

(Refer Slide Time: 08:13)



How do we do statistical analysis for this? Well, what are we looking for? Number one, we want to see that all laboratories perform the same. All the laboratory performance are same, in other

words, statistically they are close. If that is the case we would like to mix all the data and see if there are any unusual extremities there, unusual observations there.

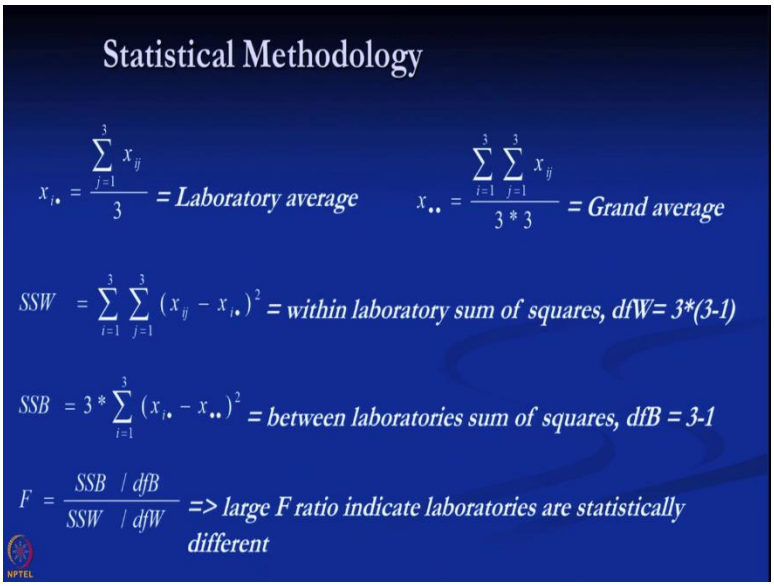
If there are no unusual observation, then we find the statistical limit within which the analysis would lie most of the time. So, we will find an interval estimate of that. Here, the emphasis is on the closeness of all the laboratories, unusual observations and most of the time, they lie in a statistical interval.

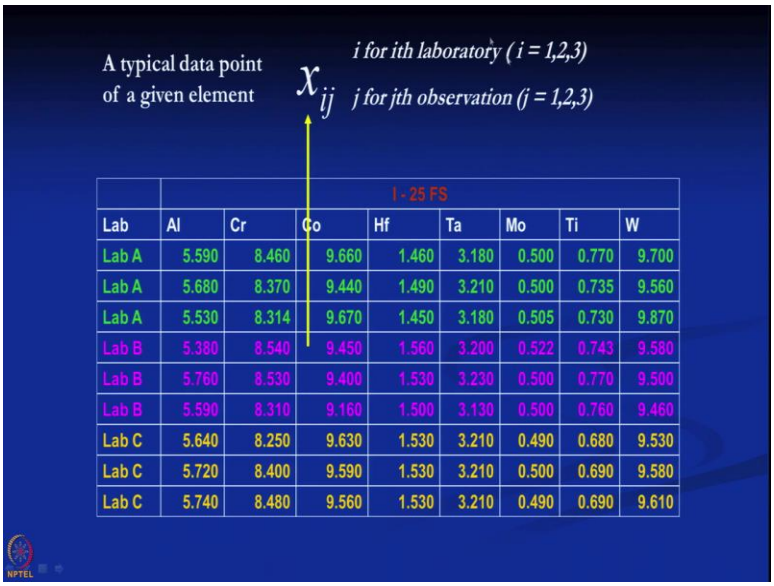
We are not going to perform the whole analysis here. We are going to look into the first part of it and that is the closeness of the, all the laboratories is what we are going to look into because that refers to analysis of variance. So, this is our data. Typically, this data is denoted as xij, where the sub-script i refers to the ith laboratory and sub-script j, sub-script j refers to the jth observation of the ith laboratory. So for example, if you look at this value 5.530, it will be called x13.

Because it is looking into the, it is the x33, it is the third laboratory looking into the third observation or you can, if you look at all the observations together, then it says that it is the third laboratory, third observation, here it will be second laboratory, 1, 2, 3, 4, 5th observation. So, likewise it is calculated.

I think I have mistaken, the first what I said is correct, it is the second laboratory, third observation. So this is x22, this is x23, this is x13, like this all the distributions are, the xij values are calculated. Let us think in a very logical way, what should be the statistical methodology? Let us define few quantities.

(Refer Slide Time: 11:23)





Let us call xi dot as an average of the laboratory. We are averaging out over the three observations of a single laboratory i. So, as in the previous case, if we take an average of these three observations, it will come to x1 dot, if you take the average of these observation it will be x2 dot. Please remember I am not taking the whole matrix of observation, I am looking at each element separately. So, let us not confuse about it.

So xi dot is the average, it is a laboratory average, x dot dot is a grand average. Then I defined, within laboratory sums of squares. So, I call it sums of squares within, SSW, which is summation

of within the laboratory. So, xij minus the laboratory average whole square. S, it is sum of squares of within the laboratory.

Because we are doing it within the laboratory and taking summation over all the three laboratory. So, this is sum of squares within laboratories. The degrees of freedom will be, this will have a degrees of freedom of 3 minus 1, because there are 3 data points and there is 1 average which has been calculated multiplied by the three different laboratories. So, it is 3 times 3 minus 1.

Then we have between laboratory sums of square. So, sum of squares between, SSB, which is 3 times the difference between the laboratory average and the grand average. So, laboratory average represents the laboratory and grand average represents the whole putting all the three laboratories together. So, this is called between laboratory sum of squares. So, it is SSB, its degree of freedom is 3 minus 1.

Now what we say is, if the variation between the laboratories and variation within the laboratories are comparable. Then you can say that there is no statistical difference between the laboratories. It means that every laboratory will have its own variation. Between the laboratories there will be a variation. If these two variations are comparable. Then we can say that there is no statistical difference between the laboratory.

So, as we say, we always talk of rejecting the hypothesis. So, we in other words, we can say that if this ratio is very large, it means that between the laboratory variation is larger than the within the laboratory variation. Then we say that these laboratories are statistically different. I think logically this argument is quite appealing. This can be written in this table.

(Refer Slide Time: 15:03)

The Result (part elements)

		Sum of Squares	df	Mean Square	F	p-value
Al	Between Labs	.026	2	.013	.864	.468
	Within Labs	.089	6	.015		
	Total	.115	8			
Cr	Between Labs	.013	2	.007	.549	.604
	Within Labs	.072	6	.012		
	Total	.085	8			
Co	Between Labs	.130	2	.065	4.627	.061
	Within Labs	.084	6	.014		
	Total	.214	8			
Ti	Between Labs	.009	2	.004	18.579	.003
	Within Labs	.001	6	.000		
	Total	.010	8			

Statistical Methodology

$x_{i\bullet} = \frac{\sum_{j=1}^{3} x_{ij}}{3}$ = Laboratory average $\qquad x_{\bullet\bullet} = \frac{\sum_{i=1}^{3}\sum_{j=1}^{3} x_{ij}}{3 * 3}$ = Grand average

$SSW = \sum_{i=1}^{3}\sum_{j=1}^{3} (x_{ij} - x_{i\bullet})^2$ = within laboratory sum of squares, dfW= 3*(3-1)

$SSB = 3 * \sum_{i=1}^{3} (x_{i\bullet} - x_{\bullet\bullet})^2$ = between laboratories sum of squares, dfB = 3-1

$F = \frac{SSB \ / \ dfB}{SSW \ / \ dfW}$ => large F ratio indicate laboratories are statistically different

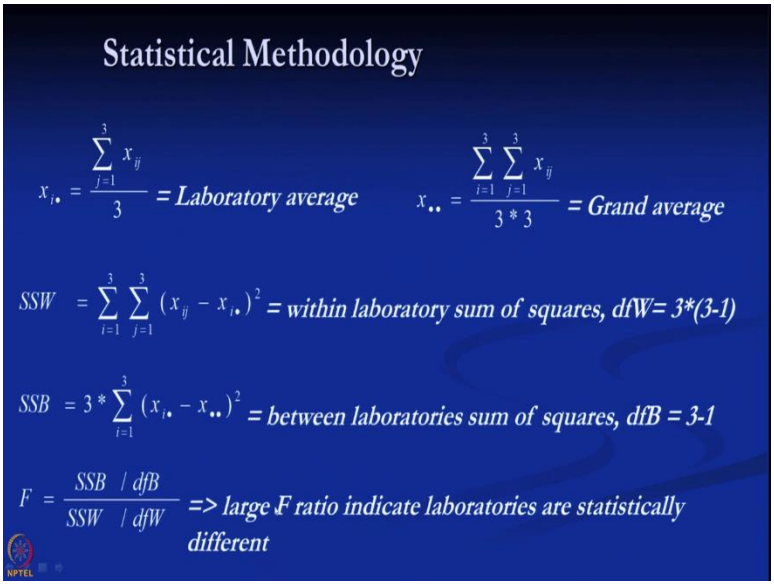
So, for example, for element aluminum, we have between labs sum of squares which is given here, its degree of freedom is 3 minus 1 which is 2, mean square is this, you please remember we have to divide it by the degrees of freedom that why we call it a mean square. Because this is sum square. So, this is mean square which is this divided by its degree of freedom.

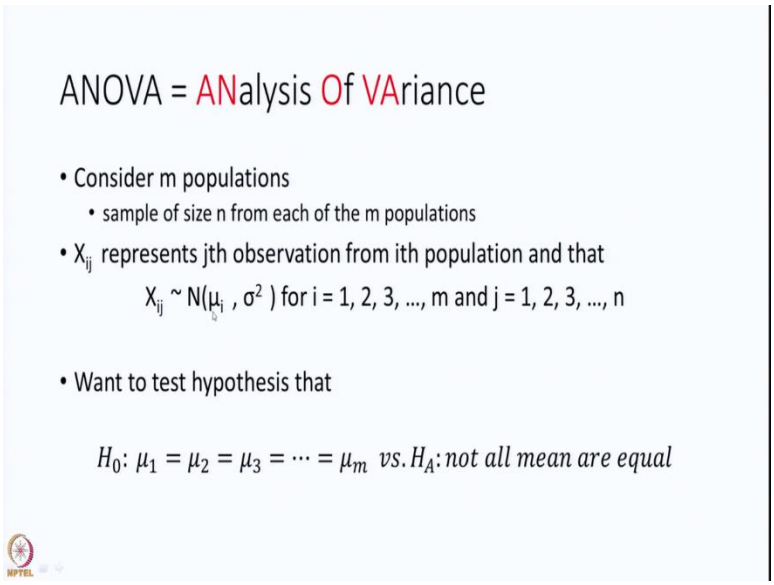
Within laboratory is so much with 6 degrees of freedom, mean square is so much. So, if you take the ratio of these two, it is 0.864, this is what I called a F value. Do not worry about p value for time being. So, likewise I have shown you typically only the partial 4 elements and I have shown

you that how the values can be larger, this is 4 point something. But you look at titanium, this has the largest value. So, you can say that here is somewhere probability, the probably the laboratories are different, at lease there is one laboratory different from the rest of them.

This is called Analysis of Variance and what you saw here for each element this table which has been made is called analysis of variance table. It is a very simple logical argument to understand that the values given by two laboratories or more than one laboratories are same or not. So, this is the, it finds the difference, the variation between laboratories, it finds the variation within laboratories and then if it finds that between the laboratory variation is larger than the within the laboratory variation. Then there is a chance that your one laboratory is, at least one laboratory is different than the rest of them. So, now let us go to the theory. This is called Analysis of Variance.

(Refer Slide Time: 17:28)

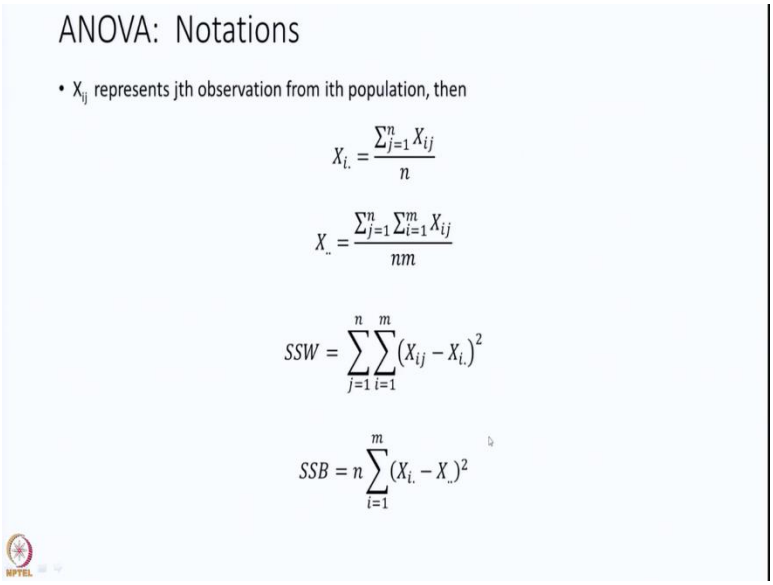




We consider m population. Now we are generalizing it. Here we had three laboratories, here we are, in this case we are going to say that we have m population and sample of size n, we had a sample of size three. Here we have a sample of size n. Then Xij represents the jth observation from the ith population and we assume for time being that Xij is normal with mean mu i and a common sigma square.

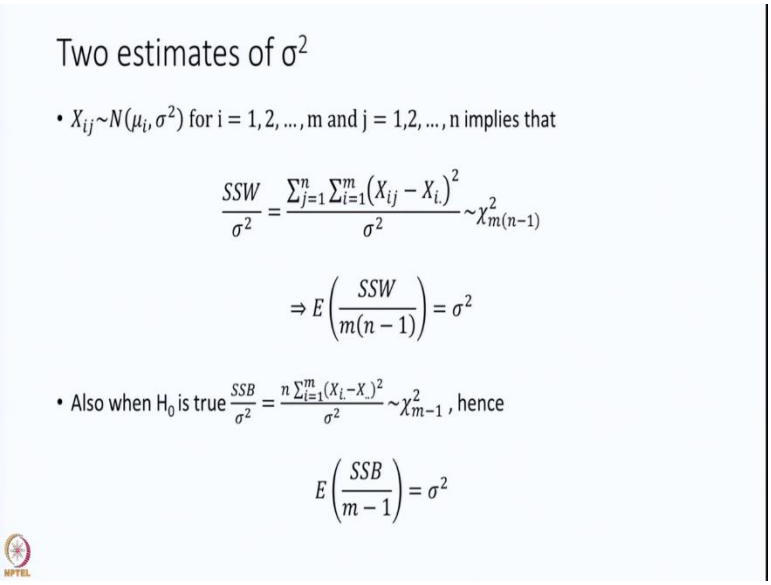
We want to test the hypothesis that all the means are same versus the alternate that all means are not same, all means are not equal. This is the hypothesis that we wish to test. Once again please remember having normal assumption, normality assumption, only is for convenience to derive the statistic distribution of F otherwise it is not really necessary. So, what I have said in the previous, if F ratio is large, then it indicates that laboratories are statistically different is enough. However, in this case we are assuming normality.

(Refer Slide Time: 19:00)



We again go through the notation Xi dot is summation j is equal to 1 to n Xij over n. This is like a laboratory average. So, this is a ith group average, ith population average. X dot dot is a grand average which is summation over j to n and i to m Xij and they are divided by n m. Then within sums of square, sum of squares. That is sum of squares within the group can be given by Xij minus Xi dot whole square and then you sum it up over i and j and between group sum of squares SSB can be given as n multiplied by summation i is equal to 1 to m Xi dot minus X dot dot whole square. This is the same procedure that we used previously.

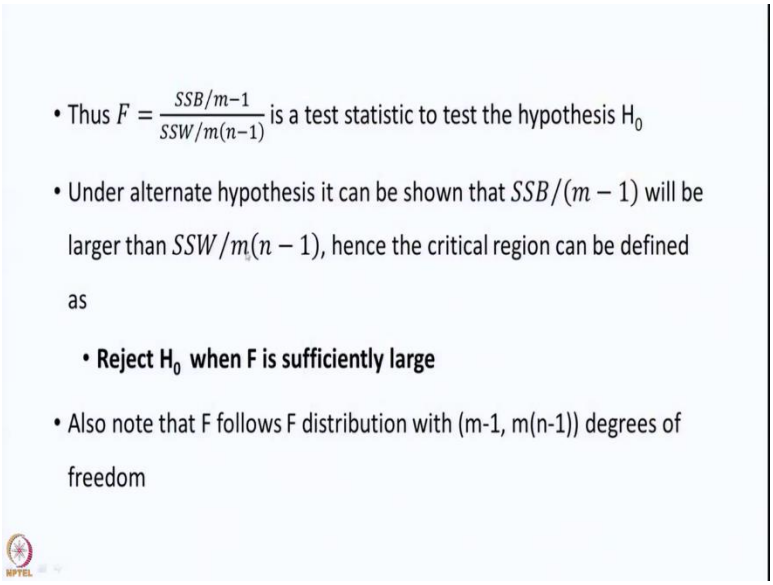
(Refer Slide Time: 20:02)



Now we have two estimates of unknown variance sigma. Xij is distributed as normal Mu plus sigma square. Then SSW divided by sigma square will be distributed as Chi square with m times n minus 1 degrees of freedom and therefore expected value of SSW by m over n minus 1 is sigma square.

Also if H naught is true, if all the means are same, then between group sums of squares is also distributed as Chi square with m minus 1 degrees of freedom and therefore expected value of sums of squares, between group sums of squares divided by m minus 1 is also sigma square. So, you see we, that is what we said in the example, the variation due to, variation between the group of, between laboratories and the variation within laboratories if you assume that all laboratories are same, they estimate the same quantity sigma square.

(Refer Slide Time: 21:23)

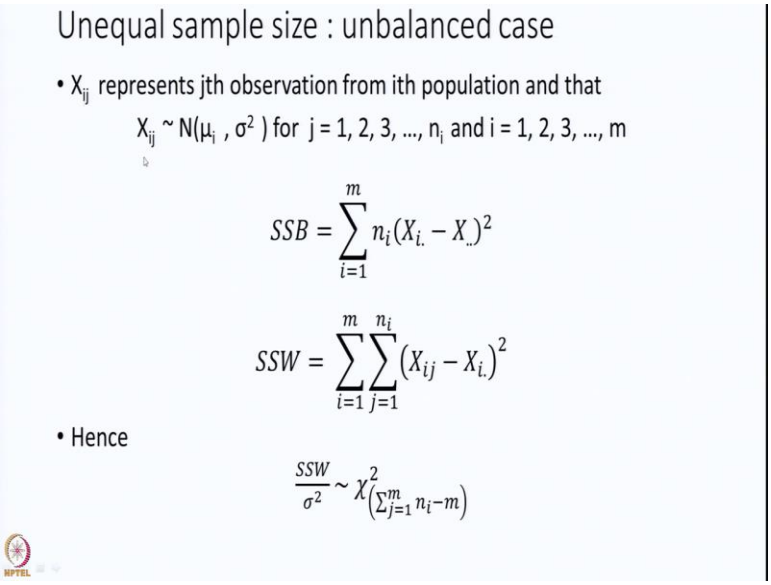


And therefore, the ratio of the two is the test statistic to test the hypothesis that all means are same. Because if hypothesis is true then this ratio has to be small. If the hypothesis is not true then the ratio is going to be larger. So, under the alternate hypothesis between groups sums of squares is going to be much larger than within group sums of squares by its degrees of freedom and therefore critical region is that reject H0 if F is sufficiently large.

If F is very large, you are going to reject the thing and remember, F is a ratio of two Chi square distribution which are independent of each other. Because all observations are taken i i, they are all taken independent independently and therefore the between group sums of squares and within group sums of squares are independent of each other.

You can show it statistically also and therefore this follows F distribution, ratio of two chi square follows F distribution with m minus 1 which is a numeral, numerator degrees of freedom divided by denominator degrees of freedom. Now you see here we have assumed that all populations are of the same size.

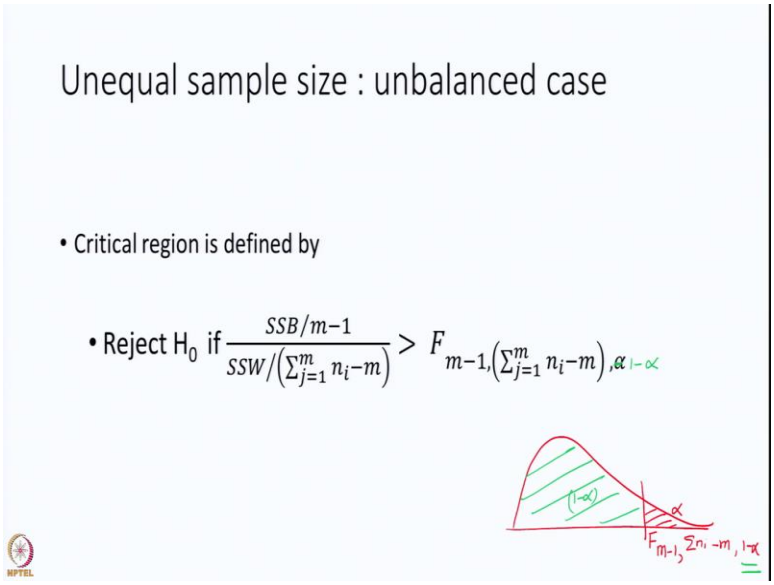
(Refer Slide Time: 22:53)



Suppose we have Xij where the i represents the population and j represents the observation and in ith observation the number of, I am sorry, ith population, the number of observation is ni. It is not say common n.

In that case only the formula changes, the between group sums of squares become summation i is equal to 1 to m ni times Xi minus Xi dot minus X dot dot square, SSW remains the same and therefore within groups sums of squares divided by sigma square is Chi square with summation of this, summation of ni minus m and between groups sums of, within groups sums of squares is not shown. But it is also a Chi square with, it will be a, it will be a Chi square with summation, it will be Chi square with m minus 1 degrees of freedom.

(Refer Slide Time: 24:20)

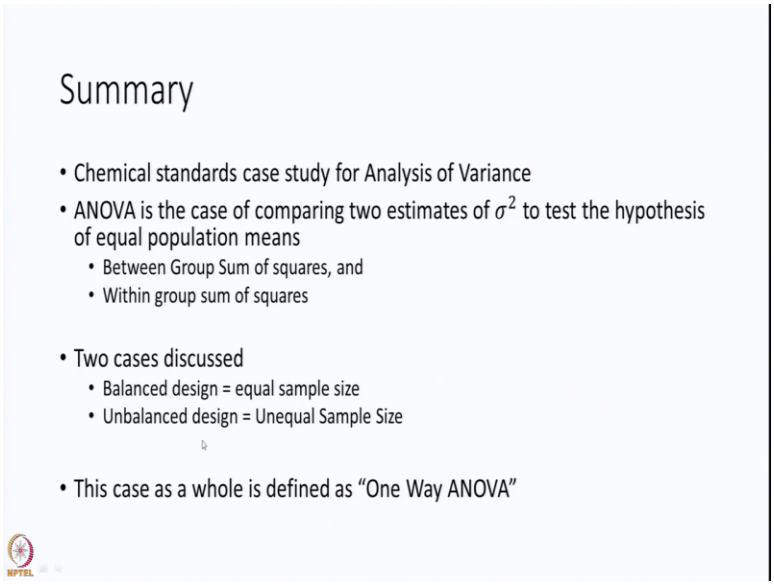


And therefore if you take the ratio, between group sums of squares divided my m minus 1, this also estimates the same population variance sigma square. Within group sums of square divided by its degree of freedom also defines the same sigma square. So, if you take the ratio, if the null hypothesis is true or then the rejection region, you reject the null, you reject the null hypothesis if this ratio is large and this large because now you assume that it is a F distribution.

You can say that it is larger than F m minus 1 and this degrees of freedom at alpha. When this happens this actually should be 1 minus alpha, let me correct myself. Because once again if you draw the graph, F is also a skewed distribution, you want to have F m minus 1 summation ni minus m and you want to have this 1 minus alpha. Because this area will be alpha and therefore, this area is going to be 1 minus alpha and therefore this value is going to be 1 minus alpha, it should be 1 minus alpha.

So, even if the, there are unequal sample size for each population, it still follows the F distribution, still we are comparing the variation between group sums of squares and within group sums of squares, we find that if all the populations are same. Then between group sums of squares and within group sums of squares divided by their appropriate degrees of freedom, the ratio should be very small. If the ratio is large, it means that between group sums of squares is large and therefore you reject the hypothesis. So, let us quickly summarize.

(Refer Slide Time: 26:44)



We started with a developing chemical standard case study, for analysis of variance. We found that analysis of variance is a case of comparing two estimates of sigma square to test the hypothesis of equal population means. They are between group sums of squares and within group sums of squares.

We discussed two cases, one is called a Balanced Design, when you have, each population has the same sample size. There is called Unbalanced Design, if the population have unequal sample size and as a whole this case is defined as one way analysis of variance. Because we have taken into account only the difference in the, in the row values only, thank you.