

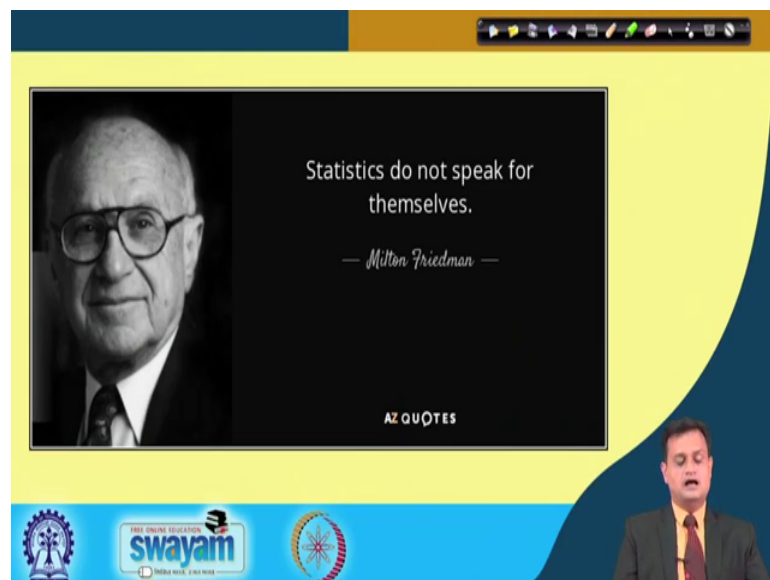
**Six Sigma**  
**Prof. Jitesh J Thakkar**  
**Department of Industrial & Systems Engineering**  
**Indian Institute of Technology, Kanpur**

**Lecture - 34**  
**One – Way ANOVA**

Hello friends, I welcome you to our ongoing Six Sigma journey. And I would like to remind you that we are in the analyze phase of our DMAIC six sigma cycle. As a part of analyze phase we have already talked about hypothesis testing, 1 tail test, 2 tail test, hypothesis testing for; 1 population, 2 population. And we have also seen the correlation and regression analysis and model validation for regression analysis.

Now, we will advance in our analyze journey and this lecture 34 will help you to appreciate a very very important concept which is widely used in design of experiment. And later on this particular concept would be used in the say next coming lectures in the design of experiment and it is One-Way ANOVA.

(Refer Slide Time: 01:21)



So, let us begin with a very good inspiring quote; 'statistics do not speak for themselves' by Milton Friedman. And these scientist Nobel Laureate says that you may manipulate with the facts the way you apply the statistics, but you need to be careful. And statistics do not speak on their own, you must try to analyze, you must try to see the wider range applicability of the results. And specifically the implications of the propose

recommendations then only your analysis, your analyze phase will have some importance.

(Refer Slide Time: 02:07)



A presentation slide titled "Recap" with a yellow background and a dark blue header. The slide lists five topics in a bulleted format, each preceded by a red square icon. At the bottom, there is a blue banner with the Swayam logo and a small inset image of a man in a suit.

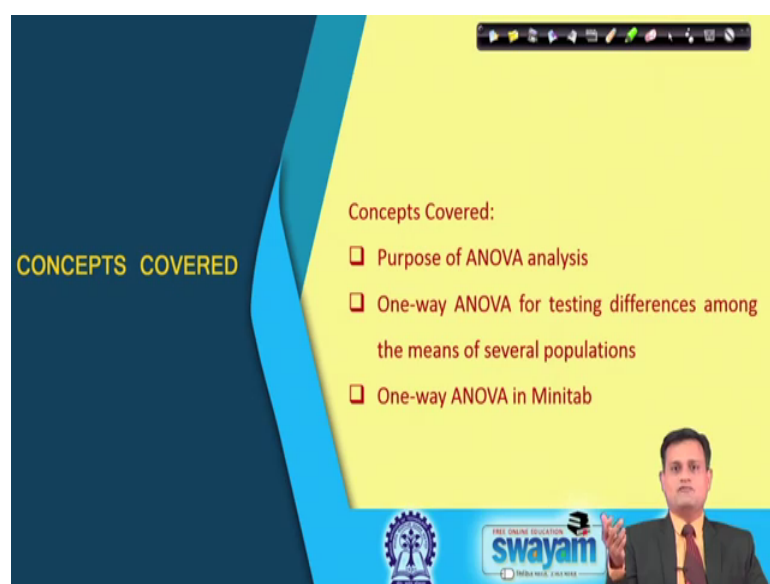
### Recap

- ❑ Validation of Regression Model
- ❑ Autocorrelation
- ❑ Durbin-Watson statistic
- ❑ The t-test and F test
- ❑ Pitfalls of regression analysis

swayam

So, if you see what we discussed in the last class then we had seen the validation of regression model. And as a part of that we had seen; autocorrelation, Durbin-Watson test, t-test, F-test, for checking the significance of coefficient  $\beta_0$  of the regression model and pitfalls of regression analysis.

(Refer Slide Time: 02:36)



A presentation slide titled "CONCEPTS COVERED" with a yellow background and a dark blue header. The slide lists three concepts in a bulleted format, each preceded by a red square icon. At the bottom, there is a blue banner with the Swayam logo and a small inset image of a man in a suit.

### CONCEPTS COVERED

Concepts Covered:

- ❑ Purpose of ANOVA analysis
- ❑ One-way ANOVA for testing differences among the means of several populations
- ❑ One-way ANOVA in Minitab

swayam

Now, in this lecture I would like to help you to appreciate the concept of ANOVA. What is the purpose of ANOVA analysis? Then we will specifically focus on the concept in example of one-way ANOVA. And then I will also try to say show some demonstration application of ANOVA, one-way ANOVA in Minitab software.

(Refer Slide Time: 03:00)

**What is 'Analysis Of Variance - ANOVA'?**

Analysis of variance (ANOVA) is an analysis tool used in statistics that splits the aggregate variability found inside a data set into two parts: systematic factors and random factors. The systematic factors have a statistical influence on the given data set, but the random factors do not.

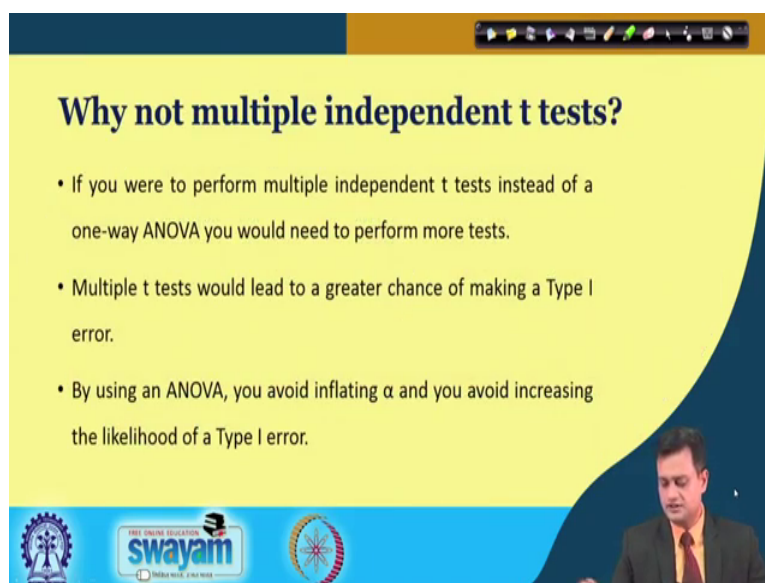
**Tells the analyst if there are any statistical differences between the means of three or more independent groups.**

So, let us see what is analysis of variance and ANOVA; if you just look at the name you will get a feel that there is something related to variance and I am trying to analyze it as simple as that. So, in a more technical term I would say that ANOVA is an analysis tool typically it is used widely used in statistical inferential analysis. And this basically splits the aggregate variability of your data set or inside a data set into 2 parts.

Number 1; systematic factor, number 2; random factor and the systematic factor basically have the influence on the given data set and random factor they do not exercise such kind of influence. So, basically this approach helps the analyst tells the analyst if there are any significant differences between the means of 3 or more independent groups.

So, I would like to draw your attention to the point that I am talking about 3 or more independent groups. And we have already seen that you can conduct the hypothesis testing for 2 population and this can help you to compare the mean of 2 different population.

(Refer Slide Time: 04:28)



### Why not multiple independent t tests?

- If you were to perform multiple independent t tests instead of a one-way ANOVA you would need to perform more tests.
- Multiple t tests would lead to a greater chance of making a Type I error.
- By using an ANOVA, you avoid inflating  $\alpha$  and you avoid increasing the likelihood of a Type I error.

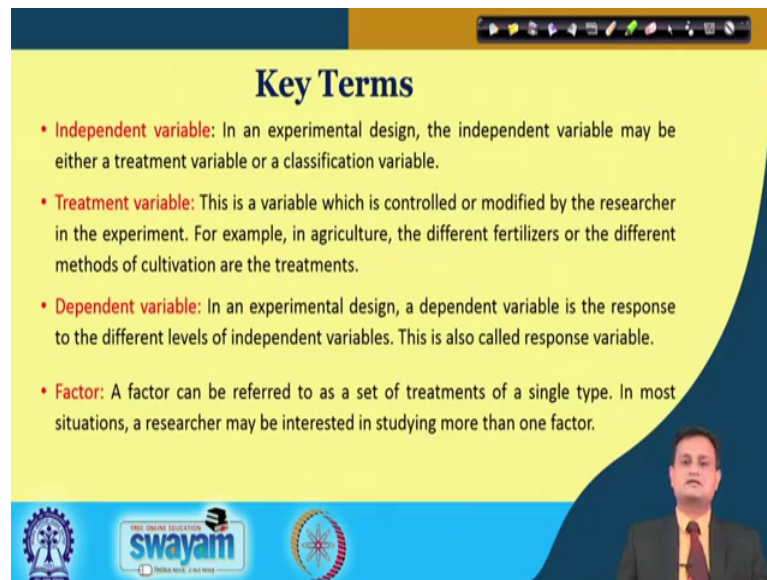
swayam  
INDIA WISE, FUTURE WISE

So, now the question here is that if I can do it for 2 for example, let us say you have to compare four mean. Now, four different population the same way as you did for 2 population, you can compare the four mean 1 by 1 each pair for C2 6 pair you can compare using the t-test or z test. Or if it is a small sample you can go for t-test. Now what is the need of this new tool.

Now, if you recall we discuss that length that when you are going for inferential analysis you always accept the risk of committing type 1 and type 2 error. So, now, here when you are going by t-test multiple t-test for comparing more than let us say 2 or 3 mean then your type 1 error it gets amplified it increases and this is where what you need you need an approach where simultaneously this different means more than 2 more than 3 can be compared.

And you can draw the conclusions based on the simultaneous analysis of your data. So, this is where the catch lies and please appreciate this fact very well that I have an option to go for multiple pair wise comparison and conduct the t test. But that will unnecessarily increase my type 1 error of making decision. And hence ANOVA analysis helps me to analyze the given situation my data simultaneously and draw the inferential conclusions.

(Refer Slide Time: 06:16)



### Key Terms

- **Independent variable:** In an experimental design, the independent variable may be either a treatment variable or a classification variable.
- **Treatment variable:** This is a variable which is controlled or modified by the researcher in the experiment. For example, in agriculture, the different fertilizers or the different methods of cultivation are the treatments.
- **Dependent variable:** In an experimental design, a dependent variable is the response to the different levels of independent variables. This is also called response variable.
- **Factor:** A factor can be referred to as a set of treatments of a single type. In most situations, a researcher may be interested in studying more than one factor.

swayam  
INDIA WISE

Now, before we actually go into the details of ANOVA analysis let us try to appreciate some of the technical terms. Number 1 independent variable, this may be typically called a treatment, variable or the classification variable. Then you have treatment variable so this is a variable which is controlled or modified by the researcher. Suppose temperature pressure, then you as a researcher you as an experimenter have a control over this factor and you can change it you can vary it.

So, for example, in agriculture the different fertilizers or the different method of cultivations are the treatments. The third one is dependent variable so in any experimental design a dependent variable is the response to the different levels of independent variable. We have seen in regression analysis that you have dependent variable and this independent variable is influenced by set of independent variables and same applies here that this is typically called the response variable.

Then you have factor so factor can be referred as a set of treatment or a single type in more situation analyst researcher may be interested in studying more than one factor. So, likewise you have couple of terms to be appreciated and this would be useful in the subsequent phase improve also. Then we will talk about the design of experiment.

(Refer Slide Time: 07:52)

## One-Way Analysis of Variance

- Evaluate the difference among the means of three or more groups

**Examples:**

- ☐ Accident rates for 1st, 2nd, and 3rd shift
- ☐ Expected mileage for five brands of tires

- Assumptions:
  - Populations are normally distributed
  - Populations have equal variances
  - Samples are randomly and independently drawn

So, let us try to appreciate the concept which we want to focus in this particular lecture, this is one way analysis of variance. And what exactly it means when I say one way because analysis of variance is clear what does it mean and what is the purpose. But when I say 1 way what does it mean.

So, here I am interested to evaluate the difference among the means of 3 or more groups. So, I am interested in a particular factor particular variable. And I am just trying to compare the mean of for that particular factor may be for more than 2 more than 3 groups. So, just try to see that accident rate for 1st, 2nd and 3rd shift.

So, how many accidents are taking place I may analyze the data, I may say collect the data for 2 years. And then I can use this data to analyze that whether the number of accidents on an average taking place in 1st shift are really different than 2nd shift and same way 3rd shift or all the 3 are same.

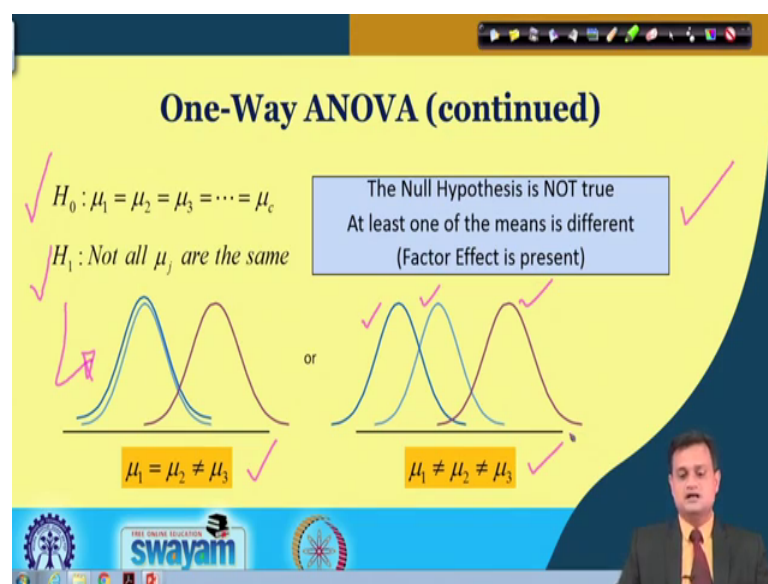
So, there are many factors as you can understand; alertness of the worker, supervise in supervision level then say kind of skills available in the different shifts that may be the case. So, this may have impact on the accident rate, but I want to analyze this fact statistically and I would like to compare the average accident rates in 3 means over a period 2 years or 3 years using the ANOVA analysis.

Similar way you can think about expected mileage for five brands of tyres. For example, you are purchasing MRF, Apollo or CEAT and many other brands you have JK tyre. Now let us say I want to see that whether there is a difference statistically significant difference in the average mileage or life of each particular brand tyre or it is more or less same.

So, again this is a fit case for conducting ANOVA analysis and analyzing the data simultaneously variability simultaneously and drawing the say a statistical conclusions. So, now, see the assumptions it is very important to remind you that statistical analysis is always based on certain assumptions. And if this assumptions are not respected or validated then your analysis will not have much value. So, here my fundamental assumption is that populations are normally distributed.

So, suppose I have  $\mu_1$   $\mu_2$   $\mu_3$   $\mu_4$   $\mu_5$  to be compared coming from different population then my all the populations they are normally distributed. Than populations have equal variances so there is a no funnel I have shown in the last lecture. There is no funnel kind of say distribution of the error and my populations are equal variances. Samples are randomly and independently drawn there is no subjectivity bias when I draw the sample for the purpose of my analysis.

(Refer Slide Time: 11:20)



Now, just see this that when I say I want to compare the mean of different coming from different population my null hypothesis says that  $\mu_1$  is equal to  $\mu_2$  is equal to  $\mu_3$



up to  $\mu_c$ . And if the null hypothesis is not true then at least 1 of the mean is different there is a factor affect available.

So, then I will say my alternate and this alternate can be represented like this at least 1 of the mean is different. So, your  $\mu_2$  is not equal to  $\mu_3$  or you may have a situation that your  $\mu_1$   $\mu_2$  and  $\mu_3$  they all 3 are not equal. So, this is how I formulate the hypothesis null and alternative for my ANOVA analysis.

(Refer Slide Time: 12:08)

**Partitioning the Variation**

- Total variation can be split into two parts:  
 $SST = SSA + SSW$  ✓

Where

- SST = Total Sum of Squares  
(Total variation) ✓
- SSA = Sum of Squares Among Groups  
(Among-group variation) ✓
- SSW = Sum of Squares Within Groups  
(Within-group variation) ✓

Now, let us try to see what is the mechanism and what exactly ANOVA analysis does. So, my ANOVA analysis typically will partition the variation into 2 part; SST is equal to SSA plus SSW. When I say SST this means total variation. So, it is basically total sum of square, I have already introduced you to the concept as well as equation of sum of square. So, SST is total sum of square total variation SSA sum of square among groups.

So, you have drawn the sample from different population. And now you have the groups available. So, I would like to compare the variability among the groups and this is my say SSA. Now, I would also like to compare or check the variability within the group. So, suppose you have drawn a sample from population 1 then this sample will have some variability within the group and this is typically called within group variation. So, please see that I have total variation I have among group variation and I have within group variation. So, these are the fundamental say concepts in my ANOVA analysis.



(Refer Slide Time: 13:37)

### Partitioning the Variation

$$SST = SSA + SSW$$

Total Variation = the aggregate variation of the individual data values across the various factor levels (SST)

Among-Group Variation = variation among the factor sample means (SSA)

Within-Group Variation = variation that exists among the data values within a particular factor level (SSW)

The slide includes a Swamyam logo and a small video inset of a man in a suit.

Now, let us try to see that what exactly we do and how we conduct the ANOVA analysis. So, we have SST is equal to SSA plus SSW I am just repeating because this is the most important concept. Total variation is the aggregate variation of the individual data value, across the various factor levels that is call SST. Among group variation this is variation among factor sample means and within group variation that exist among the data within a particular factor level or group this is call SSW.

(Refer Slide Time: 14:20)

### Total Sum of Squares

$$SST = SSA + SSW$$
$$SST = \sum_{j=1}^c \sum_{i=1}^{n_j} (X_{ij} - \bar{X})^2$$

Where

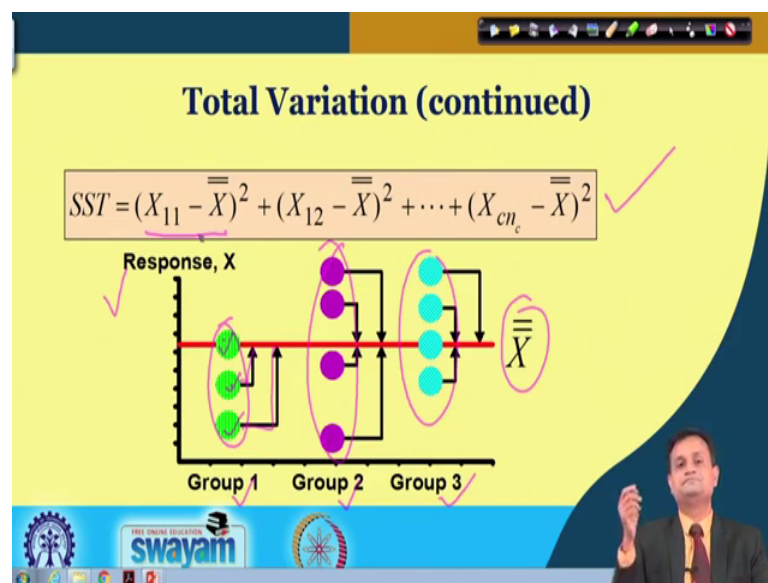
$SST$  = Total sum of squares  
 $c$  = number of groups or levels  
 $n_j$  = number of observation in group  $j$   
 $X_{ij}$  =  $i^{th}$  observation from group  $j$   
 $\bar{X}$  = Grand mean (mean of all data values)

The slide includes a Swamyam logo and a small video inset of a man in a suit.

Now, let us try to appreciate the little mathematics and how we can express the SST, SSA, SSW; so that we can compute this various sum of squares and then conduct the inferential analysis using ANOVA. So, SST is basically  $\sum_{j=1}^c \sum_{i=1}^{n_j} (X_{ij} - \bar{X})^2$ . So, SST is total sum of square  $c$  is the number of groups or levels that is why you will see that the first this thing is  $\sum_{j=1}^c$  and  $j$  is number of observations in a group that is  $j$ .

So, this is the second term and as I have  $ij$ ; obviously, I have to define  $i$  and  $j$  both. So,  $X_{ij}$  is the  $i$ th observation from group  $j$ . So, suppose you have group number 1 group number 2 group number 3 and suppose I say that 1 3 it means the first observation in group number 3 that is the  $j$ th group  $\bar{X}$  is my grand mean. So, these are the basic terms we try to define in my SST.

(Refer Slide Time: 15:41)



Now, just try to visualize so that will give you a very good idea and interest in understanding the SST. So, I compute my SST like this, I will just expanded my expression that  $X_{11}$  minus  $\bar{X}$  whole square plus  $X_{12}$  minus  $\bar{X}$  whole square plus dash dash dash plus  $X_{cn_c}$  minus  $\bar{X}$  whole square.

So, now if you see the picture, if you see the overall situation you have response  $X$ . And let us say you have group 1 you have group 2 and you have group 3. So, each particular group will have some data value each particular group will have some data set. And this is say my overall say representation we for which I would like to conduct the ANOVA

analysis. So, here you have  $\bar{X}$ . And if you see the expression let us try to take this  $\sum_{j=1}^c n_j (\bar{X}_j - \bar{X})^2$ . So, I have reading 1 2 and 3 in group 1.

So, what is the difference between these and these, these and these, these and these this is exactly on the line with respect to  $\bar{X}$  that will say become my  $\sum_{j=1}^c n_j (\bar{X}_j - \bar{X})^2$ . Similar way  $\sum_{j=1}^c n_j (\bar{X}_j - \bar{X})^2$  whole square then you will take the second group third group and each individual observation with respect to grand mean  $\bar{X}$  will be compared squared in order to find the total sum of square that is SST.

(Refer Slide Time: 17:27)

**Among-Group Variation**

$$SST = SSA + SSW$$

$$SSA = \sum_{j=1}^c n_j (\bar{X}_j - \bar{X})^2$$

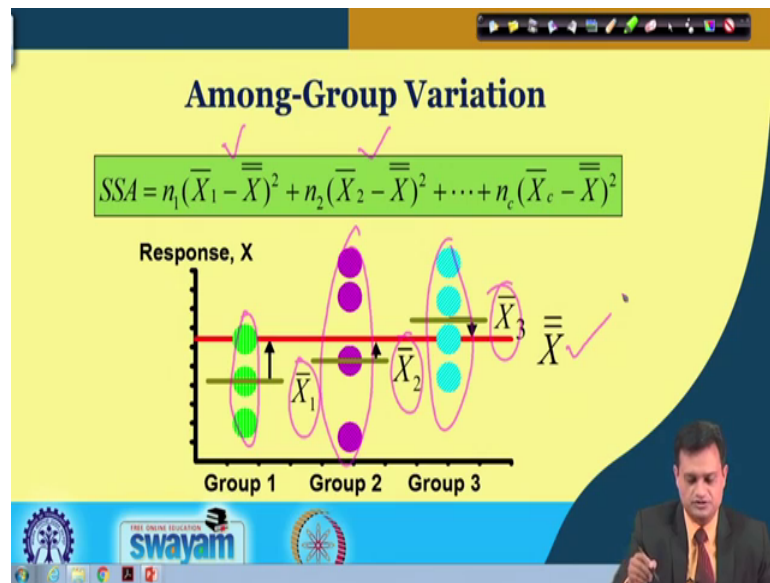
Where

- $SSA$  = Sum of squares among groups ✓
- $c$  = number of groups or levels ✓
- $n_j$  = Sample size from group  $j$  ✓
- $\bar{X}_j$  = Sample mean from group  $j$  ✓
- $\bar{X}$  = Grand mean (mean of all data values) ✓

So, with this understanding you must have appreciated what is SST. Now let us come to SSA. So, SSA is expressed as  $\sum_{j=1}^c n_j (\bar{X}_j - \bar{X})^2$ . So, SSA is sum of square among group, A stands for among. So, I have group 1 group 2 group 3 and I want to compare the variability among the group.

So, it is sum of square among the groups  $c$  is the number of groups or levels  $n_j$  is sample size from the group  $j$   $\bar{X}_j$  is sample mean from group  $j$  and  $\bar{X}$  double bar as you know it is the grand mean, mean of all the data values in the say analysis.

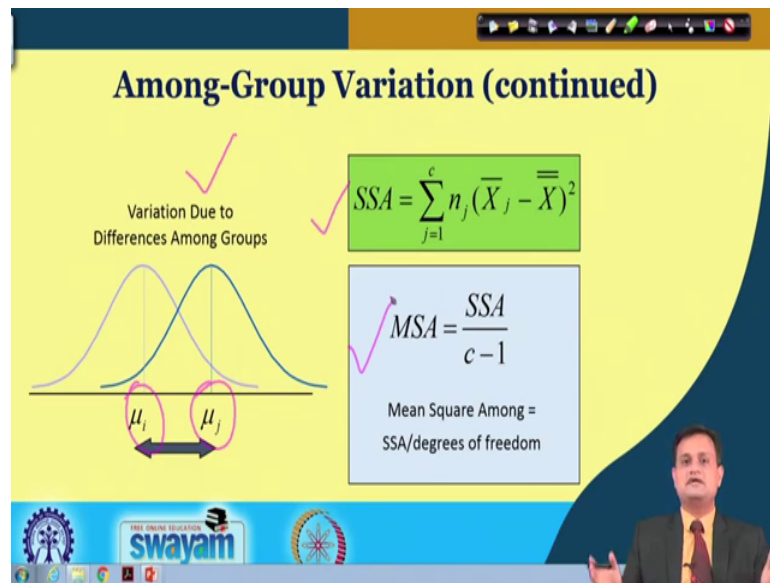
(Refer Slide Time: 18:18)



Ah so also let us try to see that how this can be visualize it would give you the better picture. So, I want to take  $\bar{X}_1 - \bar{X}$  square  $\bar{X}_2 - \bar{X}$  square. So, very simple for each particular group you find the average value that is the  $\bar{X}_1$  for group 2 I have  $\bar{X}_2$  for group 3 I have  $\bar{X}_3$ .

And take the difference with respect to  $\bar{X}$  square it multiplied with respective say sample size group size because each group may have different data point number of say units and you get the SSA, that is sum of square among the group or variability among the group.

(Refer Slide Time: 19:09)



So, now let us try to see that how can I find another term which is derived from SSA and it is basically mean square among group MSA. So, I can refer it like this that variation due to difference among group. So, you have  $\mu_i$  and  $\mu_j$ . So, there is a difference and I am capturing this difference through SSA variability.

And I can find another term, derived term that is MSA and the standard expression for MSA is say your sum of square divided by degree of freedom. So, this applies in general for whatever MSA or other you want to find, MSA is equal to SSA divided by  $c$  minus 1 in general it is sum of square divided by degree of freedom.

(Refer Slide Time: 20:10)

**Within-Group Variation**

$$SST = SSA + SSW$$
$$SSW = \sum_{j=1}^c \sum_{i=1}^{n_j} (X_{ij} - \bar{X}_j)^2$$

Where

$SSW$  = Sum of squares within groups  
 $c$  = number of groups or levels  
 $n_j$  = Sample size from group  $j$   
 $\bar{X}_j$  = Sample mean from group  $j$   
 $X_{ij}$  =  $i^{th}$  observation in group  $j$

So, we have now MSA and I have the expression which is like this, that is SSW so we have already seen SST, SSA. And now the term which is left is SSW. So, SSW stands for variability within the group I have let us say 5 data point in a group or 10 data point in a group. Now within this group among the data point what is the variability. So, I would like to see it with respect to  $\bar{X}_j$ .

So, here your  $X_{ij}$  is basically individual observation and you have  $\bar{X}_j$  this  $\bar{X}_j$  is sample mean from group  $j$ . So, for each group I will find the mean value called  $\bar{X}_j$  and each particular observation  $X_{ij}$  will be say checked or will be say the difference will be seen with respect to  $\bar{X}_j$  I will square it so I get SSW. So, this is the third term in my ANOVA analysis.

(Refer Slide Time: 21:16)

The slide is titled "Within-Group Variation (continued)". It features a yellow background with a blue header and footer. The header contains a toolbar with various icons. The main content area includes:

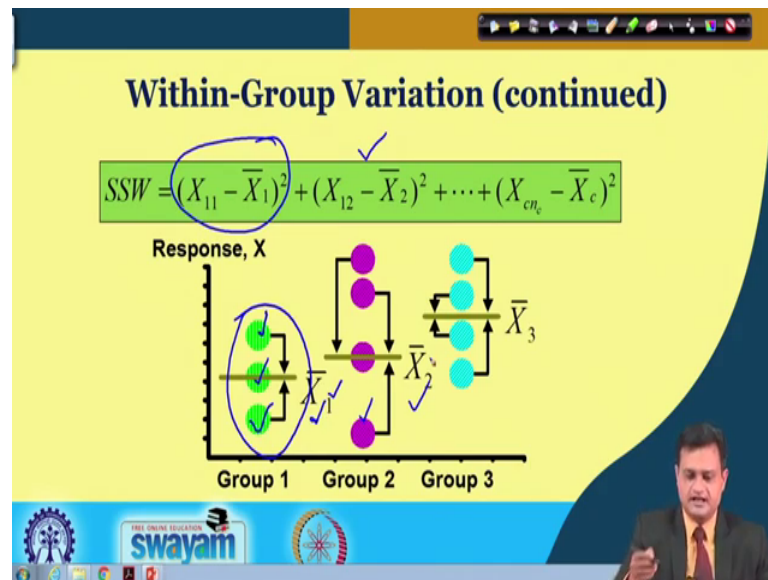
- A text box on the left: "Summing the variation within each group and then adding over all groups".
- A normal distribution curve with a horizontal line at the mean  $\mu_j$  and a double-headed arrow indicating the spread. A pink checkmark is next to it.
- A green box containing the formula for Sum of Squares Within (SSW):
$$SSW = \sum_{j=1}^c \sum_{i=1}^{n_j} (X_{ij} - \bar{X}_j)^2$$
- A blue box containing the formula for Mean Square Within (MSW):
$$MSW = \frac{SSW}{n - c}$$
- Below the MSW formula, it says "Mean Square Within = SSW/degrees of freedom".
- Handwritten pink notes on the right side: "Total DF = DF<sub>among</sub> + DF<sub>within</sub>".
- A small video inset in the bottom right corner shows a man in a suit and tie speaking.
- The footer contains logos for "swayam" and "FREE ONLINE EDUCATION".

And I would once again like to say see it that what does it mean. So, it means summing the variation within each group and then adding the overall groups. So, SSW is this and I can divide a SSW by degree of freedom  $n$  minus  $c$ . So, you will get say mean square within and this will be my expression for MSW. So, just to help you can find this very easily you have total degree of freedom is equal to degree of freedom associated with among as well as degree of freedom associated with within.

So, you have degree of freedom for total degree of freedom among degree of freedom within just try to put the values of total degree of freedom. Then you already know the degree of freedom say among we have taken it say in the previous slide. Then you will get the degree of freedom within so this is how you can easily say find the value of degree of freedom that is  $n$  minus  $c$  here within.



(Refer Slide Time: 22:32)



Now, just try to appreciate how does it look like when I say SSW. So, SSW you can see that I want to take the difference between  $X_{11}$  minus  $\bar{X}_1$  whole square this means that I have the data point in a particular group. And  $X_{11}$  minus  $\bar{X}_1$  means this particular reading minus  $\bar{X}_1$ , then  $X_{12}$  it means the group 1 second reading  $X_{12}$  I will have with respect to  $\bar{X}_1$  sorry  $\bar{X}_2$   $X_{31}$   $\bar{X}_1$ .

Similar way you can take it for this that  $X_{12}$  minus  $\bar{X}_2$   $X_{12}$  minus  $\bar{X}_2$  and so on. So, I would like to take the difference between the individual reading within a particular group with respect to say mean of that particular group and this will give me variability within the group typically called as SSW.

(Refer Slide Time: 23:41)

### Obtaining the Mean Squares

The Mean Squares are obtained by dividing the various sum of squares by their associated degrees of freedom

$$MSA = \frac{SSA}{c-1}; \text{ Mean Square Among (Degree of freedom} = c-1)$$

$$MSW = \frac{SSW}{n-c}; \text{ Mean Square Within (Degree of freedom} = n-c)$$

$$MST = \frac{SST}{n-1}; \text{ Mean Square Total (Degree of freedom} = n-1)$$

$T = A + W$   
 $n-1 = c-1 + n-c$   
 $n-c = n-1$

So, finally, I have this equations as a summary MSA is equal to SSA divided by c minus 1, MSW is equal to SSW divided by n minus c MST is equal to SST divided by n minus 1. And as I mentioned that n minus 1 is the total c minus 1 is among within, if you just equate total is equal to among plus within.

So, n minus 1 is total is equal to among is c minus 1 plus within you will get the degree of freedom n minus c this will get cancel out. So, n minus c for MSW so this is very simple to find the degree of freedom.

(Refer Slide Time: 24:36)

### One-Way ANOVA Table

Source of Variation	Degrees of Freedom	Sum Of Squares	Mean Square (Variance)	F
Among Groups	c - 1	SSA	$MSA = \frac{SSA}{c-1}$	$F_{STAT} = \frac{MSA}{MSW}$  c = number of groups n = sum of the sample sizes from all groups
Within Groups	n - c	SSW	$MSW = \frac{SSW}{n-c}$	
Total	n - 1	SST		

Now, finally, my ANOVA table will look like this please understand that we always expressed final results of ANOVA analysis in a table and this table whatever book you may refer more or less it is standard also when you say analyze your data using any software the pattern of this particular table is standard and it looks like this that you have source of variation. So, there could be variability among group, there could be variability within group and there could be total variability you have degree of freedom for each one you have sum of square among sum of square within sum of square total you can divide this by degree of freedom and you get the MSA and MSW.

Now, you are interested to analyze the hypothesis using the ANOVA analysis for more than 2 or 3 population. So, you would say that whether the mean coming from 2 to 3 population are same or different. So, obviously I need to have the statistics to check my claim at a given level of alpha. So, here my statistic is F statistic MSA divided by MSW. And I can just take the ratio of this two find the calculated value of F STAT compare it with the critical value the rest of the procedure for hypothesis testing remain same as we discussed in the previous lectures.

(Refer Slide Time: 26:15)

**One-Way ANOVA F Test Statistic**

$$H_0: \mu_1 = \mu_2 = \dots = \mu_c$$

$$H_1: \text{At least two population means are different}$$

• Test statistic

$$F_{STAT} = \frac{MSA}{MSW}$$

MSA is mean squares among groups

MSW is mean squares within groups

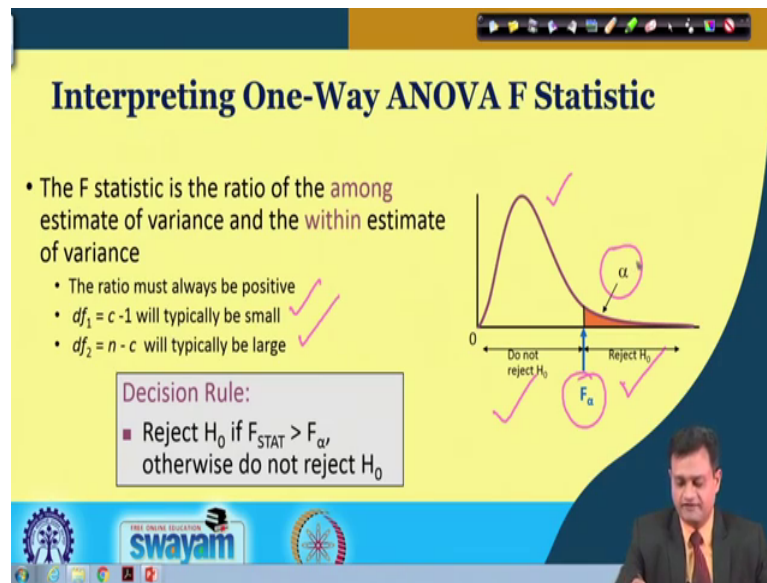
✓ Degrees of freedom

- ✓  $df_1 = c - 1$  (c = number of groups)
- ✓  $df_2 = n - c$  (n = sum of sample sizes from all populations)

swayam

So, let us try to appreciate with the help of some example set basically I want to do this. I want to compare the mean of say more than 2 or 3 population and I want to say that at least two population means are different. So, F statistic is MSA divided by MSW this is the degree of freedom.

(Refer Slide Time: 26:35)



Now, rule is very simple I have the alpha level of significance maybe 0.05 0.1, you have the  $F_{\alpha}$  which is the critical value obtained from the given level of alpha from the table. And if it falls in this region you reject the null hypothesis it means you say that your means are different at least 1 of the mean is different and here you do not reject the null hypothesis.

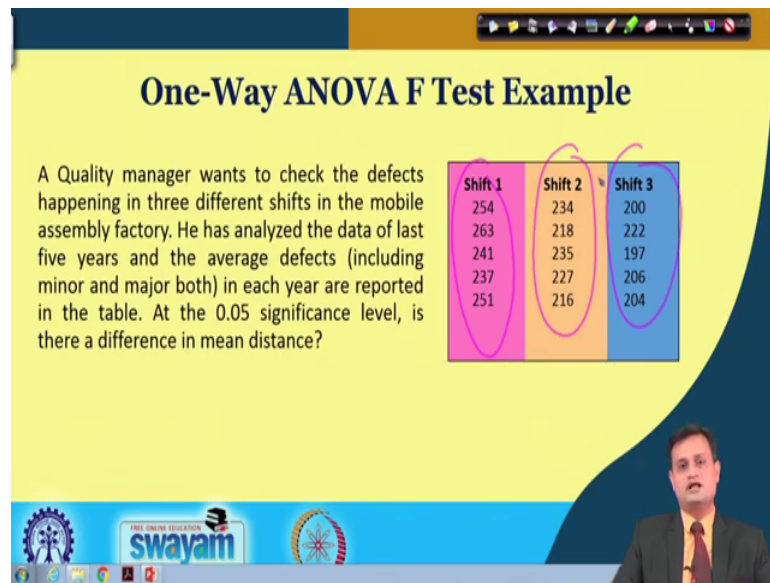
So, this is typically my f distribution which if you have to refer the table it will required degree of freedom 1 degree of freedom 2 and then you can find the f value critical value for the given level of alpha.

(Refer Slide Time: 27:24)

### One-Way ANOVA F Test Example

A Quality manager wants to check the defects happening in three different shifts in the mobile assembly factory. He has analyzed the data of last five years and the average defects (including minor and major both) in each year are reported in the table. At the 0.05 significance level, is there a difference in mean distance?

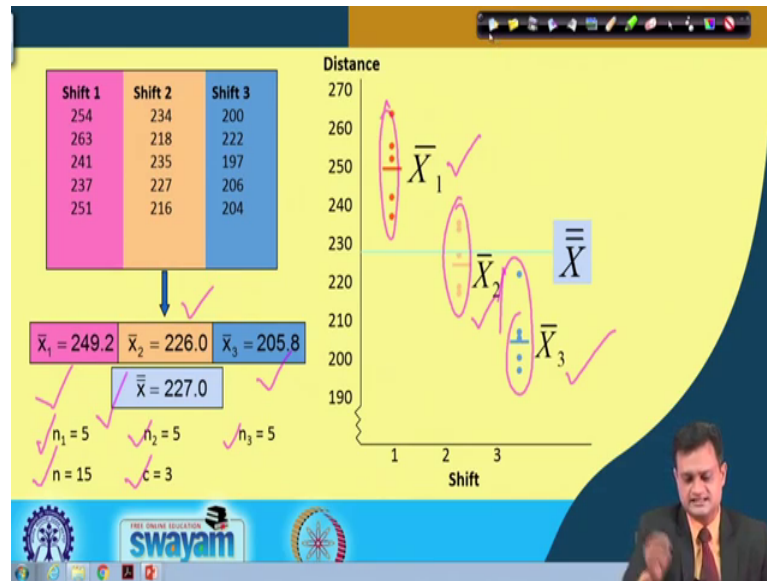
Shift 1	Shift 2	Shift 3
254	234	200
263	218	222
241	235	197
237	227	206
251	216	204



Now, let us try to see the example that will make the idea very clear. So, the example is like this you are a quality manager and you want to see the level of quality number of defectives produced in shift 1, shift 2 and shift 3 you have collected the data over a period of 5 years.

And let us say you have taken particular data for shift 1, shift 2 and shift 3 and you want to analyze based on this that whether number of defective or the defectives in shift 1  $\mu_1$  is equal to shift 2  $\mu_2$  is equal to shift 3  $\mu_3$  or they are not equal. So, this is what I want to check using the ANOVA analysis.

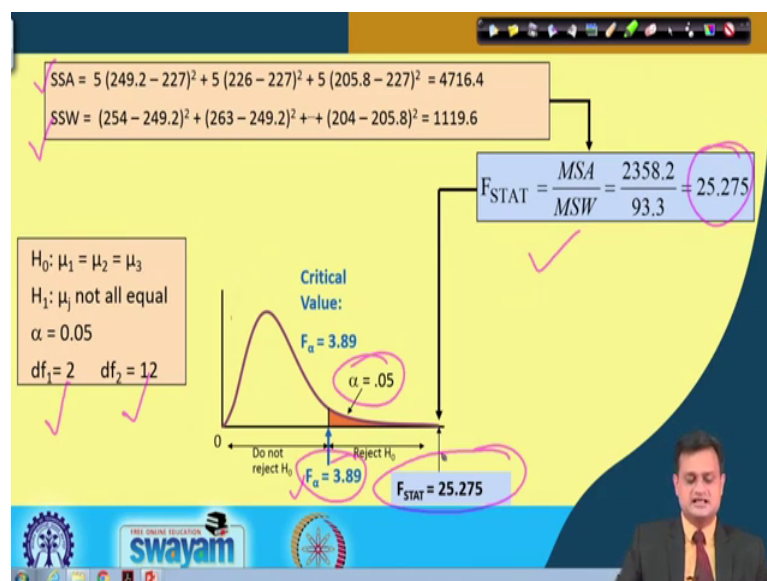
(Refer Slide Time: 28:13)



So, let us try to see the solution I will compute  $\bar{x}_1$ ,  $\bar{x}_2$ , then  $\bar{x}_3$  and I have  $\bar{\bar{x}}$  this is  $n_1 = 5$  and  $n = 15$   $c = 3$ .

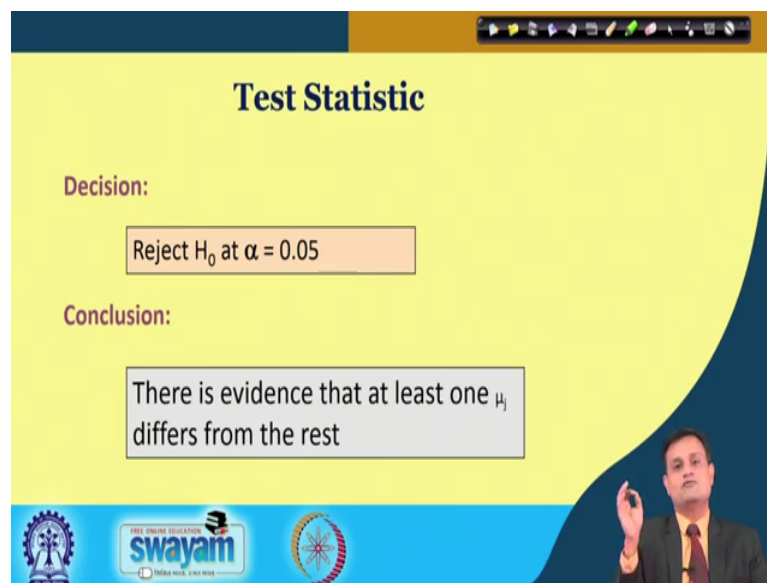
So, now you can very well understand what does I mean. So, this is my  $\bar{x}_1$  this is my  $\bar{x}_2$  this is my  $\bar{x}_3$ . So, here 3 different shifts can be seen as 3 different groups. And I have the data set in each particular group means for each particular shift, this I want to analyze simultaneously using my ANOVA analysis.

(Refer Slide Time: 28:51)



And you can see that I have computed SSA using the expressions we have already discussed SSW this is my ratio 2 find the calculated value of the F statistic and this comes out to be 25.275. Now I am selecting or general you go by alpha is equal to 0.05 the corresponding value is F alpha from the table for given degree of freedoms is now you have df 1 and df 2. So, for alpha is equal to 0.05 and df 1 is equal 2, df 2 is equal to 12 I have F alpha is equal to 3.89. So, this will make the think very clear crystal clear my F STAT is greater than this. So, this is in the rejection region.

(Refer Slide Time: 29:49)



**Test Statistic**

**Decision:**

Reject  $H_0$  at  $\alpha = 0.05$

**Conclusion:**

There is evidence that at least one  $\mu_j$  differs from the rest




And what I would say that reject null hypothesis at alpha 0.05 and there is an evidence that at least 1  $\mu_j$  differs from the rest. It means; yes, statistically there is a significant evidence to say that defectives produced in shift 1, shift 2 and shift 3 are not equal there are some causes there are some reasons that one of the shift has maybe higher or lower defectives.



(Refer Slide Time: 30:20)

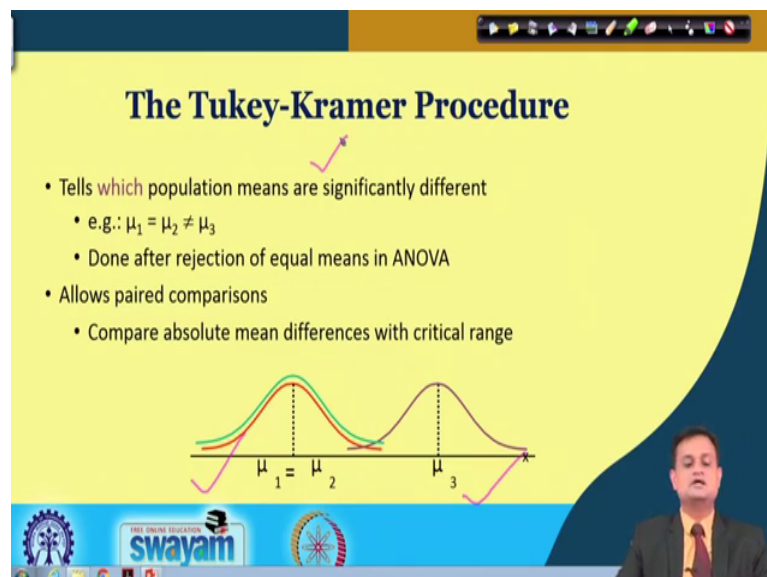
### One-Way ANOVA Excel Output

SUMMARY					
Groups	Count	Sum	Average	Variance	
Shift 1	5	1246	249.2	108.2	
Shift 2	5	1130	226	77.5	
Shift 3	5	1029	205.8	94.2	
ANOVA					
Source of Variation	SS	df	MS	F	P-value
Between Groups	4716.4	2	2358.2	25.275	4.99E-05
Within Groups	1119.6	12	93.3		
Total	5836.0	14			

So, this is how you can apply ANOVA you can also see the output in excel we will see also in Minitab more or less we will find similar kind of table constituting sources of variation degree of freedom sum of square mean sum of square and the computed value of F statistic.

(Refer Slide Time: 30:37)



Now, let us try to see the another test which basically helps me to little bit go into detail of my ANOVA analysis. When I say null hypothesis is rejected at a given level of significance it means at least one of the mean for the given level of significance is not

equal; now this does not end the story. I want to really figure out that whether the example we have discussed whether it is shift 1 and 2 or shift 1 and 3 or shift 2 and 3 yes I want to check that which particular two means are not really equal. Then the situation is described like this, that just see here;  $\mu_1$  and  $\mu_2$  they are matching, but  $\mu_3$  is not matching. It could be otherwise  $\mu_1$  and  $\mu_3$  are matching  $\mu_2$  is not matching I want to test it using some say scientific procedure and this procedure is called Tukey-Kramer procedure.

(Refer Slide Time: 31:44)

**Tukey-Kramer Critical Range**

$$\text{Critical Range} = Q_{\alpha} \sqrt{\frac{MSW}{2} \left( \frac{1}{n_j} + \frac{1}{n_{j'}} \right)}$$

where:

- $Q_{\alpha}$  = Upper Tail Critical Value from Studentized Range Distribution with  $c$  and  $n - c$  degrees of freedom (see appendix "Percentage Points of Studentized Range Statistics")
- $MSW$  = Mean Square Within
- $n_j$  and  $n_{j'}$  = Sample sizes from groups  $j$  and  $j'$

(Note: Kindly refer the book "Design and Analysis of Experiments" for appendix "Percentage Points of Studentized Range Statistics")

So, let us try to see for the example we are discussing. And what I will do I have the standardize my expression for critical range which is used for Tukey Kramer test and what I need to find I have MSW I know  $n_j$ , I know  $n_{j'}$ . So,  $n_j$  and  $n_{j'}$  sample size is from group  $j$  and  $j'$ . Here you have shift 1 2 and 3.

Now you need to have the value of  $Q_{\alpha}$ . So, this value of  $Q_{\alpha}$  is basically upper tail critical value from studentized range distribution with  $c$  and  $n - c$  degree of freedom. So, when you will refer the suggested textbook you will find this kind of table in the appendix and you can take the value of  $Q_{\alpha}$  for the given level of significance.

(Refer Slide Time: 32:37)

1. Compute absolute mean differences:

$$|\bar{x}_1 - \bar{x}_2| = |249.2 - 226.0| = 23.2$$

$$|\bar{x}_1 - \bar{x}_3| = |249.2 - 205.8| = 43.4$$

$$|\bar{x}_2 - \bar{x}_3| = |226.0 - 205.8| = 20.2$$

2. Find the  $Q_\alpha$  value from the table in appendix Percentage Points of Studentized Range Statistics with  $c = 3$  and  $(n - c) = (15 - 3) = 12$  degrees of freedom:

$$Q_\alpha = 3.77$$

Shift 1	Shift 2	Shift 3
254	234	200
263	218	222
241	235	197
237	227	206
251	216	204

$\bar{x}_1 = 249.2$   $\bar{x}_2 = 226.0$   $\bar{x}_3 = 205.8$

$\bar{\bar{x}} = 227.0$

$n = 15$   $c = 3$

So, here I can apply these for my example shift 1, shift 2 and shift 3. And what I have found that  $\bar{x}_1 - \bar{x}_2$ ,  $\bar{x}_2 - \bar{x}_3$ ,  $\bar{x}_1 - \bar{x}_3$  these are basically the differences in this mean of each particular group. And I have obtained the value of  $Q_\alpha$  3.77 for  $c$  is equal to 3 because I have 3 shifts to compare. And  $n$  minus  $c$  that is 12 degree of freedom it is 3.77. Now this 3.77 will act as a critical value comparison value. I will compare each particular difference with  $Q_\alpha$  and I will try to see whether it is greater than or it is less.

(Refer Slide Time: 33:21)

3. Compute Critical Range:

$$\text{Critical Range} = Q_\alpha \sqrt{\frac{MSW}{2} \left( \frac{1}{n_j} + \frac{1}{n_j} \right)} = 3.77 \sqrt{\frac{93.3}{2} \left( \frac{1}{5} + \frac{1}{5} \right)} = 16.285$$

4. Compare the absolute mean differences with the critical range

5. All of the absolute mean differences are greater than critical range. Therefore there is a significant difference between each pair of means at 5% level of significance.

$|\bar{x}_1 - \bar{x}_2| = 23.2$   
 $|\bar{x}_1 - \bar{x}_3| = 43.4$   
 $|\bar{x}_2 - \bar{x}_3| = 20.2$

Then and it let us say this particular difference of I am putting this  $Q_{\alpha}$  I can just go back. I am putting this  $Q_{\alpha}$  I will not directly compare I will put this  $k_{\alpha}$  into the expression we have discussed for Tukey Kramer test and his expression is basically this. So, I am putting the value of  $Q_{\alpha}$  here. So, this will give me critical range.

So, what I am getting is 16.285 now I will compare each difference with 16.285 and here all the absolute main differences are greater than critical range. Therefore, there is a significant difference between each pair of mean at 5 percent level of significance. You can say that defectives in shift 1 defectives in shift 2 and defectives in shift 3 are not equal. And say there is some evidence that there is a difference in the say defect rate per shift.

(Refer Slide Time: 34:28)

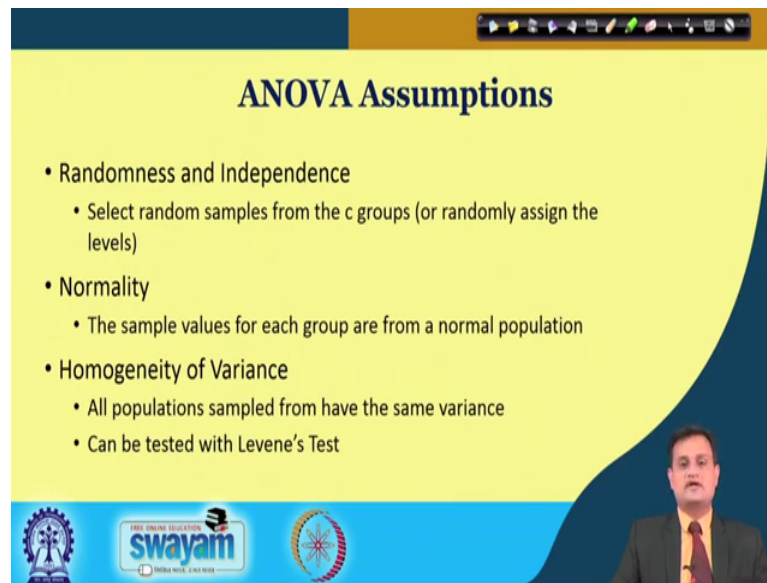
**Conclusion**

With 95% confidence we can conclude that the number of defects in Shift 1 is greater than Shift 2 and 3, and in Shift 2 - it is greater than Shift 3.

swayam

So, with 95 percent confidence we can conclude that number of defect 1 shift, in shift 1 is greater than shift 2 and 3 in shift 2 it is greater than shift 3 and so on.

(Refer Slide Time: 34:40)



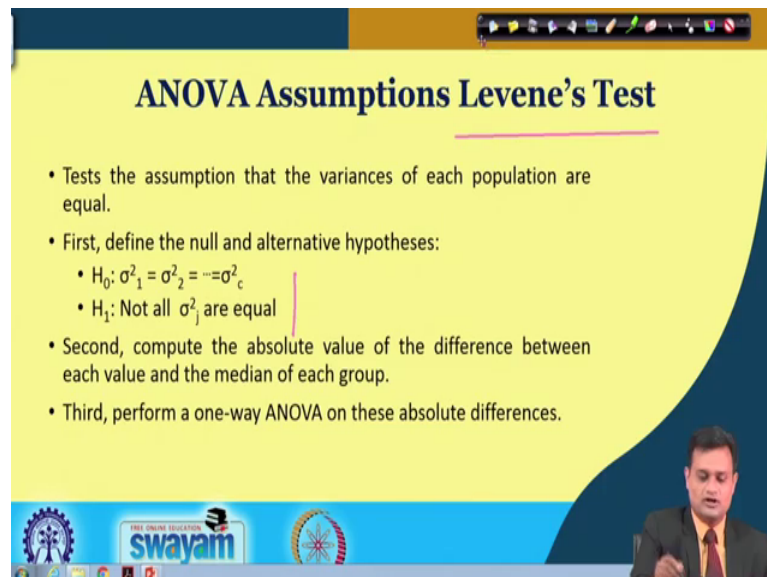
### ANOVA Assumptions

- Randomness and Independence
  - Select random samples from the  $c$  groups (or randomly assign the levels)
- Normality
  - The sample values for each group are from a normal population
- Homogeneity of Variance
  - All populations sampled from have the same variance
  - Can be tested with Levene's Test

The slide features a yellow background with a blue wavy border on the right. At the bottom, there are logos for 'swayam' and other educational institutions, along with a small video feed of the presenter.

So, there are certain ANOVA assumptions that need to be checked we will quickly try to appreciate this. So, randomness and independence we have done such kind of analysis in the previous lecture that is correlation and regression normality homogeneity of the variance.

(Refer Slide Time: 34:56)



### ANOVA Assumptions Levene's Test

- Tests the assumption that the variances of each population are equal.
- First, define the null and alternative hypotheses:
  - $H_0: \sigma^2_1 = \sigma^2_2 = \dots = \sigma^2_c$
  - $H_1$ : Not all  $\sigma^2_j$  are equal
- Second, compute the absolute value of the difference between each value and the median of each group.
- Third, perform a one-way ANOVA on these absolute differences.

The slide features a yellow background with a blue wavy border on the right. At the bottom, there are logos for 'swayam' and other educational institutions, along with a small video feed of the presenter.

So, my first assumption is about variances are equal. I can apply the Levene's test and check that whether my variances are equal or not. So, my null hypothesis and alternate

hypothesis for Levene's tests says that null is  $\sigma_1^2$  is equal to  $\sigma_2^2$  is equal to  $\sigma_3^2$ . And if not then it is my alternate hypothesis.

(Refer Slide Time: 35:26)

### Levene Homogeneity Of Variance Test Example

Calculate Medians

Shift 1	Shift 2	Shift 3	
237	216	197	
241	218	200	
251	227	204	Median
254	234	206	
263	235	222	

Calculate Absolute Differences

Shift 1	Shift 2	Shift 3
14	11	7
10	9	4
0	0	0
3	7	2
12	8	18

So, let us try to see for the same example I am checking the same data shift 1 shift 2 shift 3 this is the median. I have some say I have just converted my data by subtracting say this particular value 251 minus 237 or each value from this particular and I am just taking the positive value.

(Refer Slide Time: 35:54)

### Anova: Single Factor

#### SUMMARY

Groups	Count	Sum	Average	Variance
Shift 1	5	39	7.8	36.2
Shift 2	5	35	7	17.5
Shift 3	5	31	6.2	50.2

Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	6.4	2	3.2	0.092	0.912	3.885
Within Groups	415.6	12	34.6			
Total	422	14				

**Conclusion**

Since the p-value is greater than 0.05 there is insufficient evidence of a difference in the variances

So, you will get the calculated absolute differences I am taking only the positive value. Now you try to see the p value specific to this analysis between group within group this is the consolidated data. And you get the p value 0.912. So, now, I think you are comfortable in interpreting p value. So, p value is greater than 0.05. So, there is insufficient evidence of a difference in the variances. So, I can say the assumption about equal variances is satisfied and whatever ANOVA analysis I have done is true.

(Refer Slide Time: 36:36)

**Example**

In IC manufacturing, a plasma etching process is widely used. An engineer is interested in investigating the relationship between the RF power setting and the etch rate. He is interested in a particular gas ( $C_2F_6$ ) and gap (0.80 cm), and wants to test four levels of RF power: 160W, 180W, 200W, and 220W. The experiment is replicated 5 times.

The slide is part of a presentation with a blue and yellow header. The bottom of the slide features logos for 'swayam' and other educational institutions. A presenter is visible in the bottom right corner of the slide frame.

So, this is the first thing that I can check, now I have one way ANOVA analysis in Minitab and you have let us say the data set like this your manufacturing the integrated circuits. And there is a plasma etching process engineer is interested in investigating the relationship between RF power setting in the each rate and interested in a particular gas  $C_2F_6$  0.08 and wants to test four level of RF power 160 180 200 and 220. So, experiment is replicated five times.



(Refer Slide Time: 37:09)

• Step 1: Insert the values of the parameter in the worksheet

Run No.	Power Level (W)	Etch Rate (mm/min)
1	160	71.9
2	180	67.9
3	220	72.5
4	160	64.2
5	180	61.0
6	200	66.1
7	160	63.0
8	200	62.9
9	160	67.0
10	160	67.5
11	180	69.0
12	180	69.3
13	200	60.0
14	220	71.5
15	180	66.5
16	200	63.7
17	160	63.9
18	200	61.0
19	220	68.5
20	220	70.8

Levels of the treatment / input factor

Corresponding values of the response variable

We input the levels of the treatment in one column (C2) and the corresponding values of the response variable in another column (C3). This type of data input is called the **stacked** case in Minitab. It is a preferred way because it allows arranging data with the corresponding run order (in column C1) so that the independence assumption can be checked in ANOVA analysis.

So, you have this data set available now you just put this data set into your newly opened Minitab project or file and you have the run number power level each rate. So, this you can insert very easily.

(Refer Slide Time: 37:22)

Power-160	Power-180	Power-200	Power-220
575	565	600	725
542	593	651	700
530	590	610	715
539	579	637	685
570	610	629	710

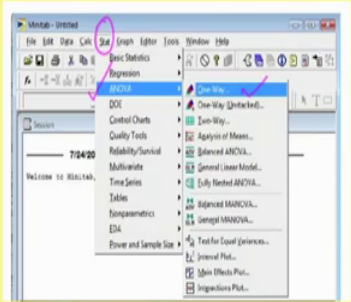
In **unstacked** case, the response values of a given treatment are inputted in a separate column. Ex: the data for Power Level 160 to 220 are stored in columns C6 through C9 respectively. Note that the Run No. cannot be inputted in unstacked case.

Then you can also put it in an unstacked manner. So, this is the unstacked manner as usual you have the data available. So, for power 160 this is one group you have this data, for power 180 group 2 you have this data power 200. So, this data is basically the etching rate and you can put it in the unstacked manner and also analyze it.

(Refer Slide Time: 37:45)

• **Step 2: Perform Data Analysis**

The example is a one-factor factorial design. To perform the One-way analysis of variance for stacked data, click Stat>ANOVA>One Way



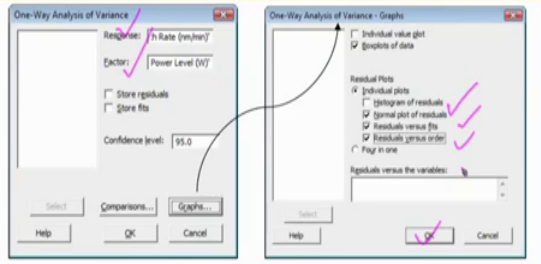
swayam

So, now the steps are very simple you to go ANOVA basically in your stat module. Go to one way and then you are try to conduct the ANOVA analysis.

(Refer Slide Time: 37:59)

In the dialogue box which appears, select "C3 Etch Rate" for **Response** and "C2 Power Level" for **Factor** by double clicking the columns on the left. Then Click **Graphs** to select the output graphs of the analysis. In the dialogue box, check "Boxplots of data", "Normal plot of residuals", "Residuals versus fits" and "Residuals versus order". Then Click **OK** back to previous dialogue box. Click **OK** again to generate the results of the One-way ANOVA.

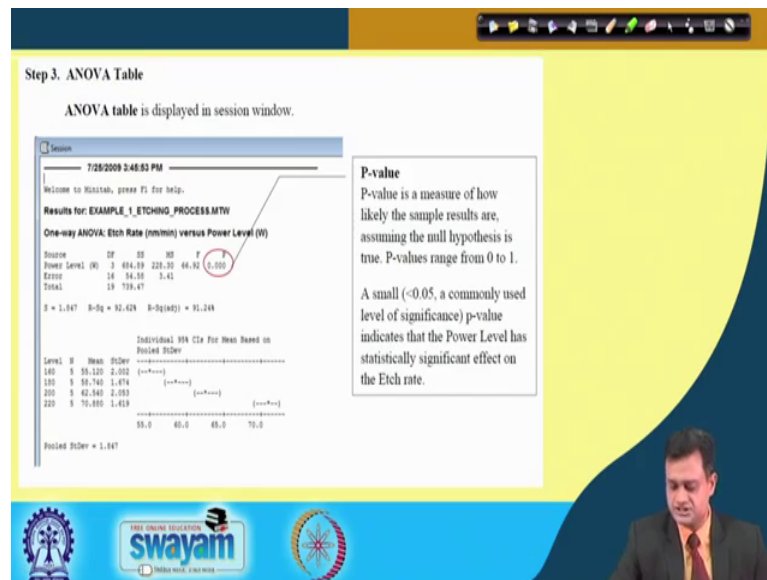
The One-way ANOVA table is displayed in the session window. The boxplot, normal plot of residuals, residuals versus fits, and residuals versus order graphs are popped-up.



swayam

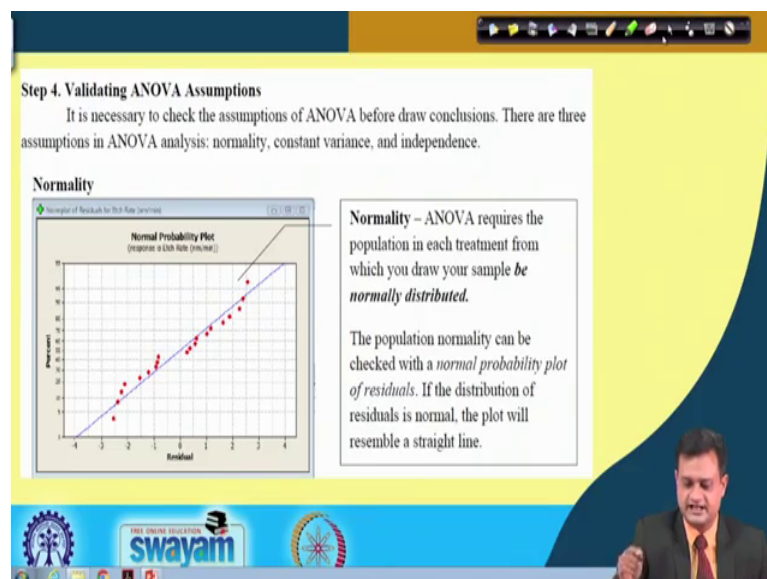
And you put response variable factor here is power level which you are changing. And also try to get all the graphs, so that you can really validate your ANOVA model press and you will have the output available.

(Refer Slide Time: 38:18)



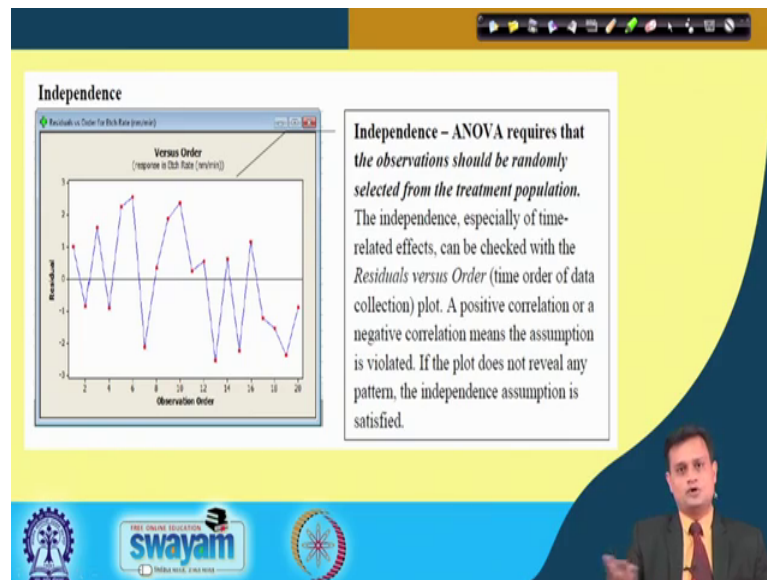
So, basically you get first thing that whether your claim is 2 or not so p is 0.00. So, this is less than 0.05. So, this says that I am in the rejection region, I reject the null hypothesis. So, etching rate for different power setting is not equal that is a difference.

(Refer Slide Time: 38:44)



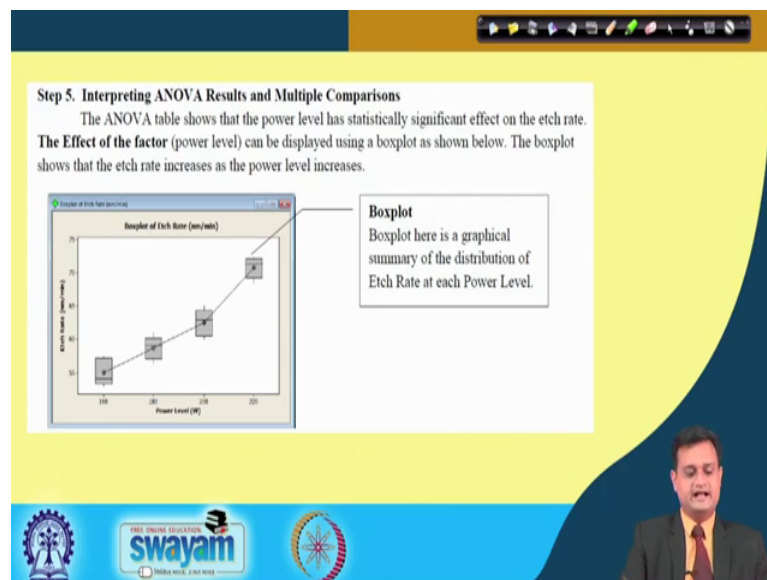
Now, you can have other outputs from Minitab like this, normality plot more or less the line which you see here is passing through the data set and I can say that my normality assumption is satisfied. So, you are putting the residual versus percentage on the normality plot.

(Refer Slide Time: 38:59)



You have the equal variance constant variance assumption you have independence no particular trained is observed.

(Refer Slide Time: 39:07)



You can also check your box plot and typically see the distribution of the etching rate.

(Refer Slide Time: 39:18)

1. ANOVA statistically compares\_\_\_\_\_.

2. What is the utility of Levene's test?

3. The null hypothesis for ANOVA is that all means are not equal. (True/False)

4. What are the key steps for conducting ANOVA analysis in Minitab?

So, this kind of plots you will get when you conduct the ANOVA analysis. Before I end as a usual practice let me float couple of think it ANOVA statistically compares, fill in the blank. What is the utility of living test null hypothesis for ANOVA is that all means are not equal true or false just revisit it. And what are the key steps for conducting ANOVA analysis the in Minitab. So, try to digest the content properly and appreciate the various concepts in detail.

(Refer Slide Time: 39:49)

**References:**

- Montgomery, D C. Design and Analysis of Experiments, Wiley.
- T. M. Kubiak, Donald W. Benbow, The Certified Six Sigma Black Belt Handbook, Pearson Publication.
- Forrest W. Breyfogle III, Implementing Six Sigma, John Wiley & Sons, INC.

Refer this references if you have any particular difficulty. And I hope the idea would be clear.

(Refer Slide Time: 39:55)

*Conclusion*

ANOVA is a statistical method used to test the differences between two or more means. It is used to test general differences rather than specific differences among means.

IIT Bombay

swayam

FREE ONLINE EDUCATION

PROGRESS WITH PURPOSE

So, as a summary I would say that ANOVA is a statistical method used to test the differences check the variability within and among and typically check that two or more mean whether they are equal or not two or more population they are equal or not.

So, it is used to test general differences rather than specific differences among the means. So, thank you very much for your interest in learning the concept of ANOVA and typically one way ANOVA. We will keep discussing the various topics as a part of our ongoing phase that is the analyze phase in DMAIC cycle. Till the time keep revising enjoy be with me.