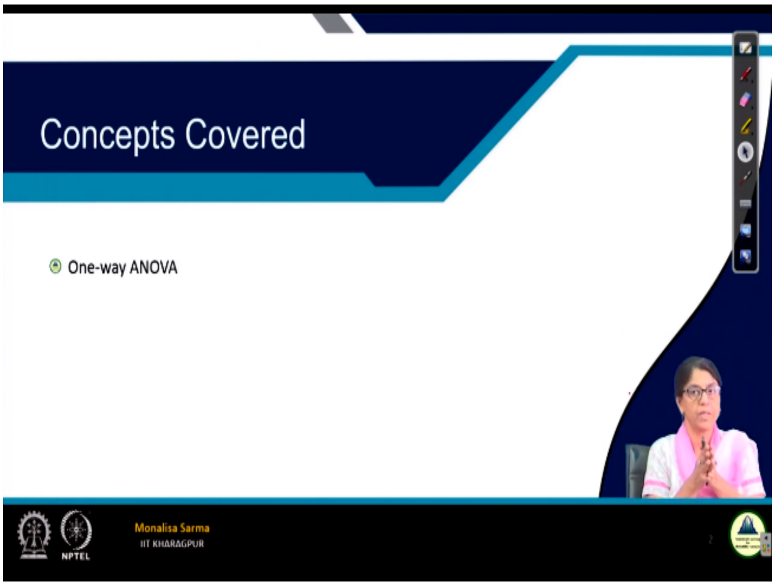
Statistical Learning for Reliability Analysis

Dr. Monalisa Sarma
Subir Chowdhury of Quality and Reliability
Indian Institute of Technology – Kharagpur

Lecture – 34
ANOVA - III

Hello guys. So, we were discussing analysis of variance in short ANOVA so in today's discussion and continuation of our earlier discussion, we have already taken 2 classes on analysis of variance today is the third class on that and continuation of that.

(Refer Slide Time: 00:43)

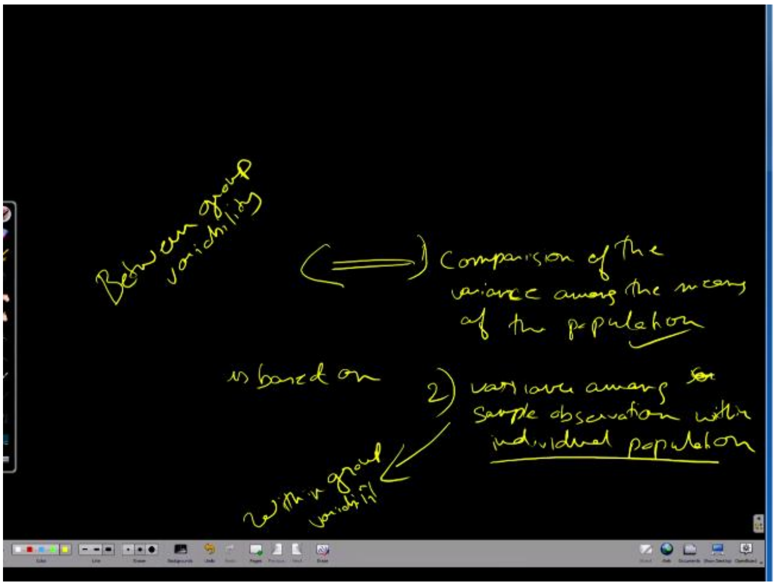


Concepts Covered

- One-way ANOVA

Today we will be discussing one way ANOVA before that, let us take a quick recap let me use the board here.

(Refer Slide Time: 00:56)



Between group variability

is based on

1) Comparison of the variance among the means of the population

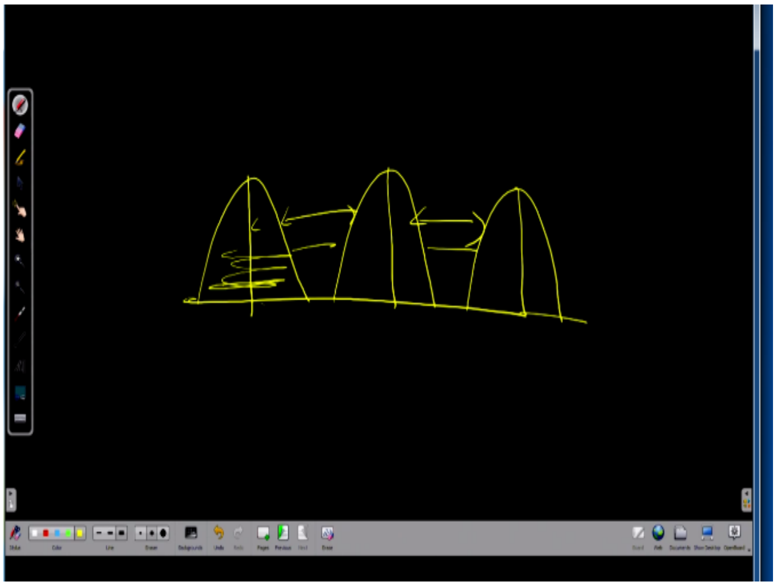
2) variance among ~~for~~ sample observation within individual population

Within group variability

So, remember what we have seen the analysis of variance is based on let me write it is basically based on 2 factors we have seen what are those 2 factors? Let me write it here one is the comparison of the variance among the means of the population first let me write it and then the second one is like variance among sample observation within individual population we have seen this analysis of variance is based on what one is the comparison of the variance among the means of the populations, this is what we call it this thing?

We call it between group variability or between treatment variability, anything this we call is between group variability and this second one that is variance among the sample observation within individual population, this is what this is called within group variability.

(Refer Slide Time: 03:06)



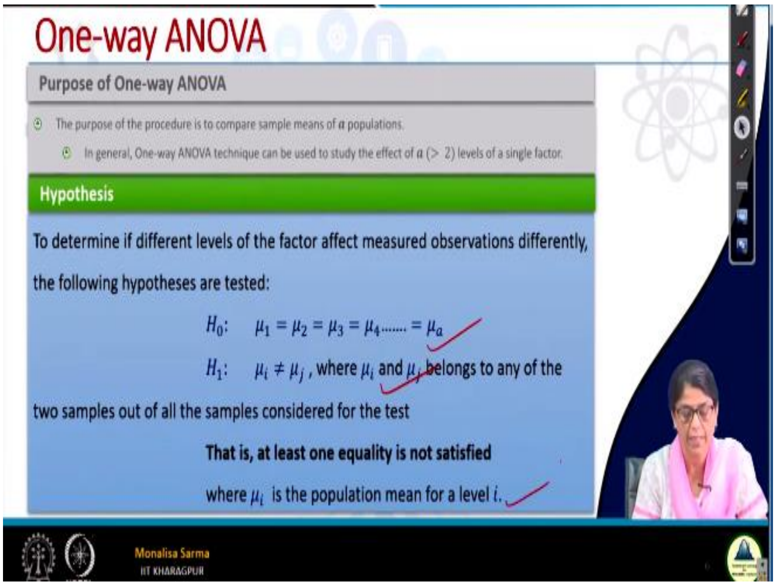
So, if we basically draw the diagram, if I draw the diagram say this is one sample this is another sample, I am giving the distribution of the sample basically this is one sample. So, this is the mean this is the mean this is the mean. So, first one is within when I tell is between group variability as between group variability means, this variability between this group variability between these 2 groups, variability between these 2 groups this is called between group variability and another is within group variability.

Within group variability this variability within the group that mean then what is the standard deviation basically, standard deviation is what is the how much it deviates from the mean each and every data how much it deviates from the mean. So, this is called within group variability and these things are called between group variability. So, we had analysis of variance is basically based on this comparison of these 2 variance, we have discussed till here and we will gradually proceed from there like. So, based on this also we have seen.

But in ANOVA what is the concept of dependent variable, what is the concept of independent variable we have also seen that. So and now, in today's class, we will be discussing one way ANOVA one way ANOVA means as I told you if you can remember, so, there is one independent variable and one dependent variable which depends on one independent variable. Like again if I can come to my previous example for the class which the performance of the student based on the music.

So, the performance of the student was the dependent variable independent variable was the music so, that is called one way ANOVA where we have both dependent and independent variables this one, so, we will be discussing today one way ANOVA.

(Refer Slide Time: 04:59)



One-way ANOVA

Purpose of One-way ANOVA

- The purpose of the procedure is to compare sample means of a populations.
  - In general, One-way ANOVA technique can be used to study the effect of a ($>$ 2) levels of a single factor.

Hypothesis

To determine if different levels of the factor affect measured observations differently, the following hypotheses are tested:

H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4 .... = \mu_a$

H_1: $\mu_i \neq \mu_j$, where μ_i and μ_j belongs to any of the two samples out of all the samples considered for the test

That is, at least one equality is not satisfied

where μ_i is the population mean for a level i.

So, now as we have already discussed and this is just a quick recap the purpose of the one way ANOVA is to compare sample means of a population, a population means when I am telling a population means that means there is only one factor since we have one way ANOVA there will be only one factor. For one factor there are how many levels if there are a levels that means total will have a populations like for the music, there are one factor that is music is factor, how many levels whether there are 3 levels like.

And even I have given an example of medicines, medicine a treatment, different types of treatment for a particular disease. So, treatment a, b, c so that was also one factor and there were 3 levels. So that was a population's means 3 populations. And again, we have given a single example of concrete mixer, the moisture observation of a concrete mixer, if we add different types of chemicals; I have used 5 different types of chemicals. So, there is 5 different types of chemicals that means total there are 5 levels.

So that means here 5 population a is 5 a is the number of level basically. In general one way ANOVA technique can be used to study the effect of a levels of a single factor where a is greater than 2, but at the same time for even if a = 2 still we can use ANOVA I will not give you the answer you can thing yourself for two populations, can we use ANOVA? If yes, why do not we use ANOVA why you use T test, if you want to compare two populations.

Very simple answer you also yourself will be able to get the answer for that. So, now, coming to that, how we go about the test, how to go about the whole thing? So, our main aim here is why do we do ANOVA, what is the reason behind? First of all we are interested in trying to find out compare the means of a population. So, basically we want to find out because of a particular treatment, is there any difference in the different populations we have given some treatment.

Because of the treatment is there any difference in among the population or there is no difference, that is what we basically want to find out. So, now, so, we will like previously for what statistical inference what we have done similarly, here also will become a carrying out hypothesis test. So, our null hypothesis is that all the μ's are equal implying that the treatment has no effect on the populations.

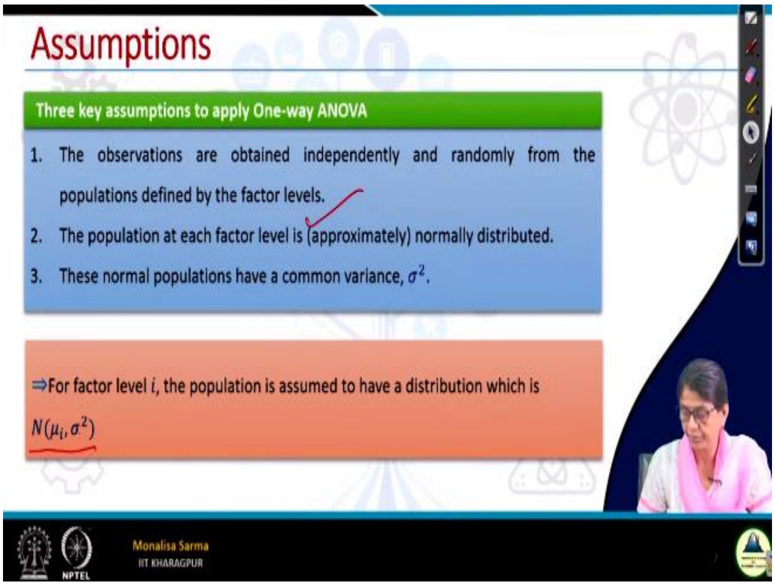
So that implies the null hypothesis, null hypothesis all the means are equal means, we have done the treatment like we have played the music, but music or no music or constant music, whatever it is the it did not have any effect on the performance of the students, it totally depends on the student's own calibre. So, that is my null hypothesis. So, for the mixing chemical to the concrete, they are also the mixing with the form.

If I mix different chemical also the moisture observation there is no change in the moisture observation that is my null hypothesis. Now my what is my alternate hypothesis, alternate hypothesis is that at least 2 of the means are not equal that means the factor is having some or the other effect because of the treatment, we are getting some effect there is a change in the mean and at least 2 of them means, there can be change in more than 2 there can be 3 are different 4 are different or 5 populations, all the 5 are different.

But at least 2 means are different that is the alter means alternate hypothesis means that is all the means are not equal, basically, that is my alternate hypothesis. Now, as I told you in my last lecture, so if I am interested in finding out among if among the between the 2 hypotheses in my alternate hypothesis gets accepted the means when my null hypothesis gets rejected and alternative hypothesis gets accepted if my alternate hypothesis gets accepted I can only say that all the means are not equal at least 2 are not equal.

But through these tests through ANOVA I am not being able to say which of the means are not equal for that different there are separate tests which I have already mentioned in my last lecture. So, now carrying with this, so, this is my null hypothesis this is my alternate hypothesis. And the way μ i is the population mean for level i.

(Refer Slide Time: 09:28)



Assumptions

Three key assumptions to apply One-way ANOVA

1. The observations are obtained independently and randomly from the populations defined by the factor levels.
2. The population at each factor level is (approximately) normally distributed.
3. These normal populations have a common variance, σ^2.

⇒For factor level i, the population is assumed to have a distribution which is $N(\mu_i, \sigma^2)$

Now, when we do one way ANOVA we have to take certain assumption, like for our different statistical inference also, you see, we have taken some assumption what was the assumption when we have taken T distribution, chi square distribution, F distribution that our population is normal. We have taken the assumption that our parent population is normal with that assumption only we have to work on a statistical inference though T test is quite robust.

But still very much away from normal is not acceptable. χ^2 and F is very much sensitive to normal. So, here also similarly we have certain assumption. So, what are the assumptions? There one of the first of them assumption is that the observations are obtained independently and randomly from the population defined by the factor levels. So, defined by the factor levels the different population from the different factor levels the different population from the levels randomly we take the samples.

There is no dependency among picking samples from here and there is no dependency randomly we pick the samples from the different factor levels populations. That is the first assumption that assumption is valid in all the tests what we have done. The population at each factor level is approximately normally distributed here also we need approximately normally distributed like T, χ^2 and F. It has one more assumption this normal population have a common variance σ^2.

So, we are trying to compare different populations a populations where a is the number of the level now, all these populations are approximately normal further this all this population has a common variance σ^2. So, these are the 3 assumption now, like remember way we have taken a common variance and while we are comparing two populations to for independent case when you are comparing two populations and the population standard deviation is not known.

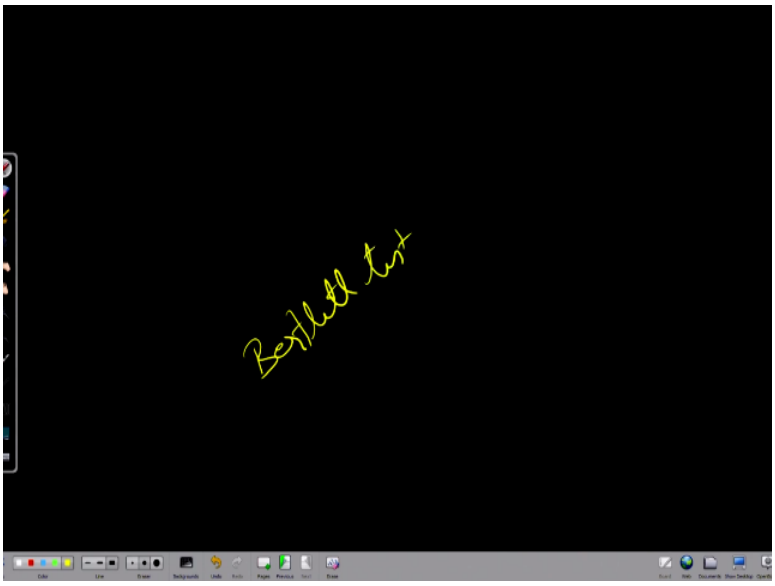
Remember then what we have taken another two populations, we are trying to compare the mean of two population and the population standard deviation is not known than what we have taken we have assumed it to be equal there is one case assumed to be equal, when we have assumed to be equal, then from there we try to find out a pooled variance estimate. Remember if you cannot remember please go to the slides again. So, from there we try to find out the pooled variance estimate what is the pooled variance estimate?

It was nothing but a weighted sum of the variance of both the populations both the samples basically weighted sum of the variance of both the sample that gives us the pooled variance estimate that it what we have already done in the while comparing two population. Similarly, here also we will assume that it has a common variance σ^2 and like in two population case, we have assumed that σ as the both the population has a common variance that is the pooled variance that we have seen.

But again they will test also 2 test whether both the variance that we have assumed If I can we assume the variance of both the populations is equal for that we have seen F test remember similarly, here also we can check because here there is no two population there will be more than two population 3 4 5 6 7 8 9 10 anything. So, we have assumed all the population has a common variance σ ^{2} and plus at the same time, here also there is a test that this assumption that it has a common variance σ ^{2} it is true or not.

So, this is beyond the scope of this course, I am not going to the test and in fact if you are interested, you can go to it yourself only it is available in the textbook and you can Google it also.

(Refer Slide Time: 13:28)



I will give the name of the test. So, it is called basically Bartlett test at this stage, you do not require basically so, I have not kept this on. Now, for our convenience and this ANOVA actually we need to be assumed that all of our population has a common variance σ ^{2}, if we do not have a common variance σ squared and can definitely we cannot use just one way ANOVA in this case, so that still if you are not sure that they may have different variance, then we can go and check whether this Bartlett's test.

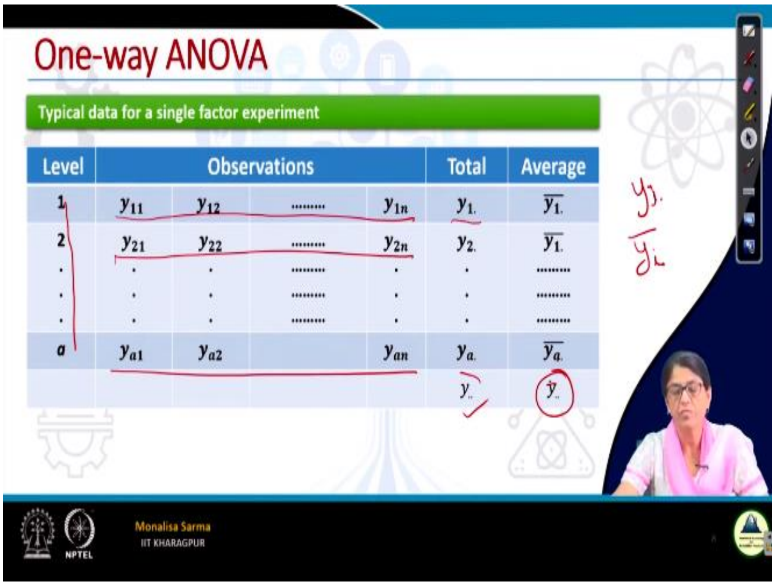
Actually, the thing is that it is very natural to assume that it will have a common variance why you see how what type of population we are not trying to compare apples and oranges. What we are doing on the same populations? We are basically doing different treatments. Like for the case of students, students score, we are basically same level of students only is not it? Either, I am talking of one level, same level of students; we are taking it from different class different section.

And what was the variance of their performance? The variance in the same level of class and student level, the variance of the performance will be same level, is not it? You are supposed to have given some treatment because of some treatment. It is not that some people will have very good result and some people will have a very bad result that is very unlikely whatever earlier what was the variance of performance before any treatment after treatment definitely their score will improve.

So, if the score improves at all the score will improve, but the variance will remain same earlier suppose before the suppose we have concluded that giving music helps the concentration level earlier suppose, a student in a class he got out of 50 his score is always around 40. So, after introduction of music maybe he got 45 another students say his performance is 35 maybe after introduction to music he got a 38 the variance basically remains the same before our level and after variance usually remains the same.

That is why it is quite intuitive to assume that the it will has a common variance that is σ ^{2}, but still if you have doubt you can go and do the Bartlett’s test. So, for factor level i the population is assumed to be have a distribution which I can mention it in this related in this way because this is what that means, it is a normal population with mean μ i and variance σ ^{2}.

(Refer Slide Time: 16:17)



One-way ANOVA

Typical data for a single factor experiment

Level	Observations				Total	Average
1	y_{11}	y_{12}	....	y_{1n}	$y_{1.}$	$\overline{y_{1.}}$
2	y_{21}	y_{22}	....	y_{2n}	$y_{2.}$	$\overline{y_{1.}}$
.	.	.	....	.	.	....
.	.	.	....	.	.	....
.	.	.	....	.	.	....
a	y_{a1}	y_{a2}		y_{an}	$y_{a.}$	$\overline{y_{a.}}$
					$y_{..}$	$\bar{y}_{..}$

$y_{i.}$

$\overline{y_{i}}$

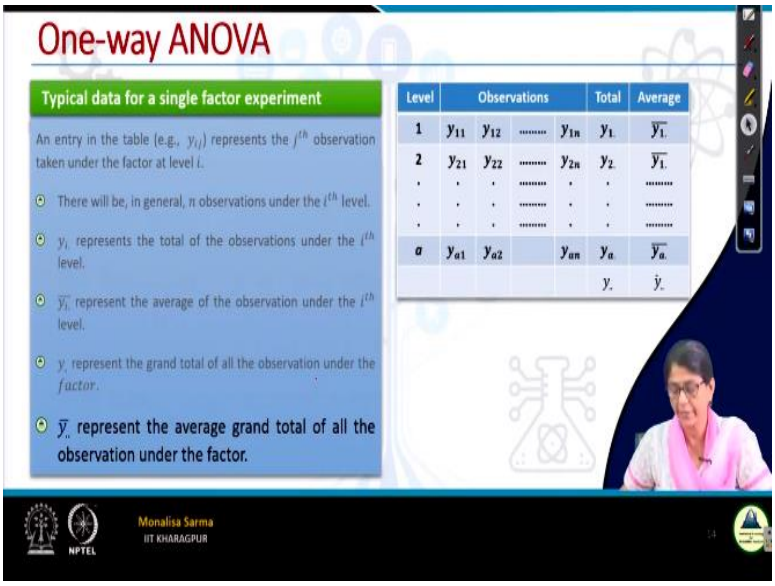
Monalisa Sarma
IIT KHARAGPUR

NPTEL

So, typical data set for a single factor experiment will look something of this sort these are the different levels we have 1 to a levels. So, this is the first data set or for level 1 the different data is for level 1, this is the different data for level 2 similarly, these are the different data for level a now, that is sum total of this I am using an expression that is y i dot see some total is y 1 dot sum total is y 2 dot that means, y i dot that not written properly y i dot then mean of this is y i dot bar.

Similarly, the grand total grand total means total of all the data how will, I get grand total this total will also give me the grand total. So, this is y double dot this one y double dot that is grand total and this is the grand mean y double dot bar is the grand mean, grand mean is the mean of all sorts all set of data this also I can get by the mean of this means, is not it?

(Refer Slide Time: 17:37)



One-way ANOVA

Typical data for a single factor experiment

An entry in the table (e.g., y_{ij}) represents the j^{th} observation taken under the factor at level i.

- There will be, in general, n observations under the i^{th} level.
- $y_{i.}$ represents the total of the observations under the i^{th} level.
- $\overline{y_{i.}}$ represent the average of the observation under the i^{th} level.
- $y_{..}$ represent the grand total of all the observation under the *factor*.
- $\bar{y}_{..}$ represent the average grand total of all the observation under the factor.

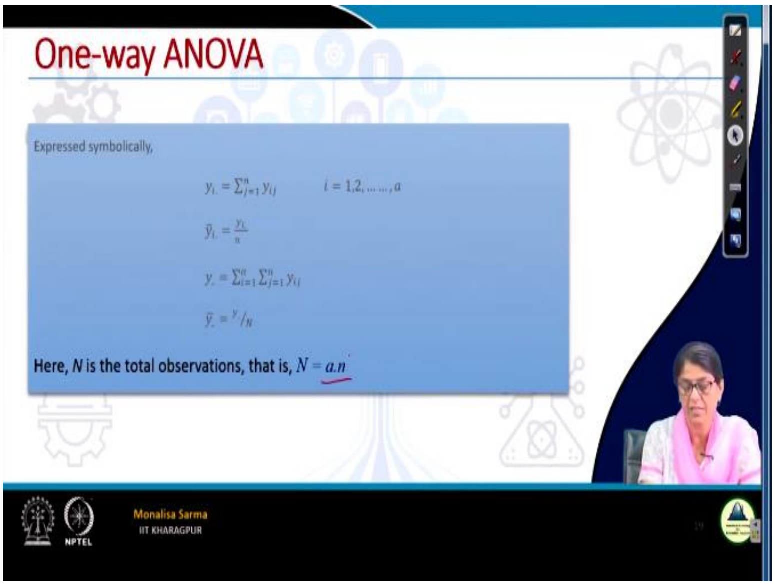
Level	Observations				Total	Average
1	y_{11}	y_{12}	....	y_{1n}	$y_{1.}$	$\overline{y_{1.}}$
2	y_{21}	y_{22}	....	y_{2n}	$y_{2.}$	$\overline{y_{1.}}$
.	.	.	....	.	.	....
.	.	.	....	.	.	....
.	.	.	....	.	.	....
a	y_{a1}	y_{a2}		y_{an}	$y_{a.}$	$\overline{y_{a.}}$
					$y_{..}$	$\bar{y}_{..}$

Monalisa Sarma
IIT KHARAGPUR

So, an entry in the table example y ij what does it represent any entry suppose y ij say y 2n. So, here i is 2, j is n. So, represent the jth observation taken under the factor level i so, this is the nth observation taken under factor level 2 so, any entry table y ij that is a representation there will be in general n observation under the ith level here first we are considering that all the sample sizes are equal we will also see when the sample sizes are not equal.

So, here we are considering that we have taken from each population we have taken an equal sample size that is n, y i dot represents a total of the observation under the ith level as I already told you, then the y i bar is the mean y double dot is the grand total and y double dot bar represents the grand mean average total means average grand mean yesterday we have come across this what grand mean remember.

(Refer Slide Time: 18:51)



One-way ANOVA

Expressed symbolically,

$$y_{i.} = \sum_{j=1}^{n} y_{ij} \qquad i = 1,2, \dots \dots, a$$

$$\bar{y}_{i.} = \frac{y_{i.}}{n}$$

$$y_{..} = \sum_{i=1}^{a} \sum_{j=1}^{n} y_{ij}$$

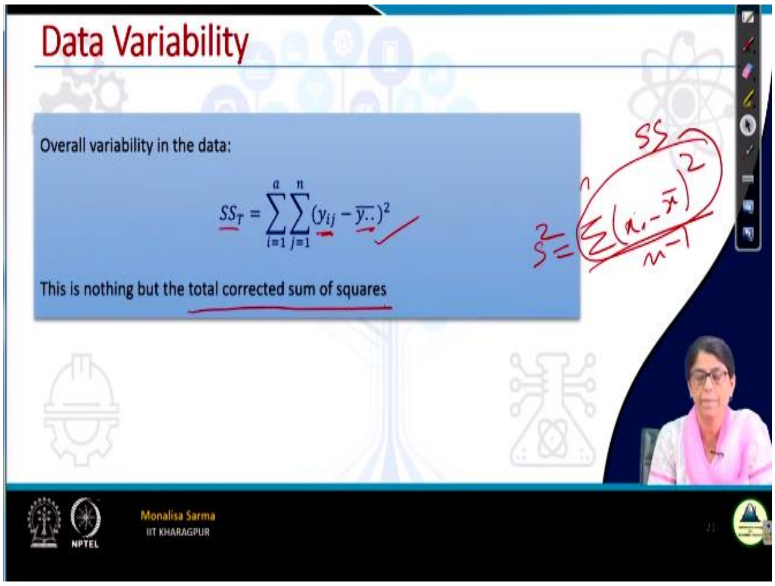
$$\bar{y}_{..} = {}^{y_{..}}/_{N}$$

Here, *N* is the total observations, that is, $N = \underline{a.n}$

So, this is expressed in this; what is y i dot, y i dot bar and all there is nothing you know like at this y i dot is the sum of all data. So, that is sum of all data basically for which will for ith level we are summing up all the jth data how many data's are the n data. So, for the ith level all the y 1 + y 2 + y 3 upto y n for the ith level that gives me my y i dot and y i dot bar is divided by n. Similarly y double dot is the summation of all data then grand mean is the y double dot by N, N is the total number of all data.

So, N is the total observation where N = a n because sample size is n total a level so, that will be N = a dot n. Now, we will see given this table we have seen this data this data this is the whole data is not it? From this whole data, now we are interested in finding out the overall data variability. We know what is variability and statistics when you are talking about statistics, variability is a term which will come across each and every step. So, you should know clearly what is that. So, now I am interested in finding out the overall data variability.

(Refer Slide Time: 20:21)



Data Variability

Overall variability in the data:

$$SS_T = \sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - \overline{y}_{..})^2$$

This is nothing but the total corrected sum of squares

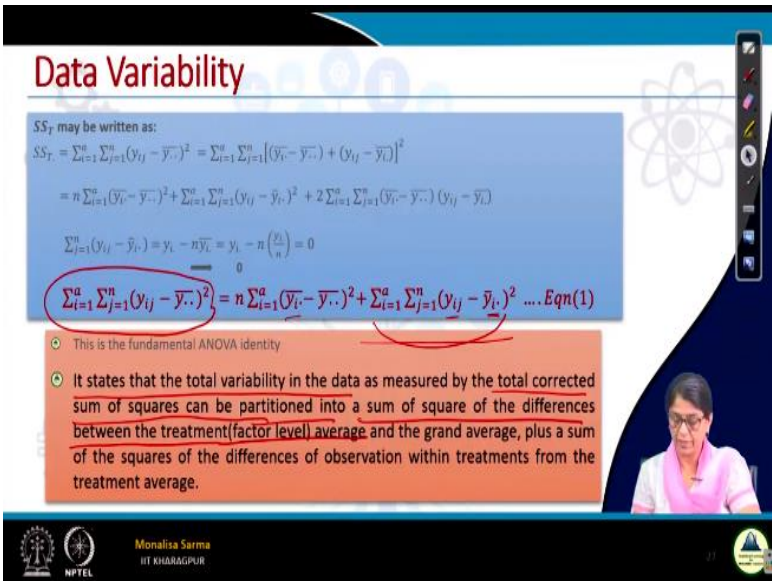
$$S^2 = \frac{\overset{SS}{\sum (x_i - \bar{x})^2}}{n-1}$$

So, overall data variability means total and let me take a small example like the music example, there are total 10 students who have taken from each class so total there are 30 students. So, first I found out the mean of all these 30 students, so that I call it grand mean. So, overall variability means how might each and every data varies from the grand mean that is made overall variability of the data, is not it?

So here, so this let me term it as an SS T, so this is each data I am subtracting it from the grand mean, this is nothing but the basically the sum of squares. Remember, what is sum of squares we have seen this is if I am interested in finding out S ^{2} for a sample, what is S ^{2}? S ^{2} is (x i - x bar)^2 this is i = 1 to n, then this is divided by n - 1. So, this factor we call it sum of squares is not it? Sum of squares SS so, this is nothing but the sum of squares only for all the levels.

So, that means, y i means we are considering all the data's all data we are subtracting from the grand mean. So, we are finding out this is the total variability of the data not from one sample from all this sample from across all the sample the whole set of data, for the music example, there are total 30 data's so, we are trying to variability from across all this 30 data by subtract each and every data from the grand mean, mean of all this data. So, that is we are calling it SS T this is nothing but a total corrected sum of squares, is not it? This is nothing but the total corrected sum of squares.

(Refer Slide Time: 22:17)



Data Variability

SS_T may be written as:

$$SS_{T.} = \sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - \overline{y_{..}})^2 = \sum_{i=1}^{a}\sum_{j=1}^{n}\left[(\overline{y_{i.}} - \overline{y_{..}}) + (y_{ij} - \overline{y_{i.}})\right]^2$$

$$= n\sum_{i=1}^{a}(\overline{y_{i.}} - \overline{y_{..}})^2 + \sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{i.})^2 + 2\sum_{i=1}^{a}\sum_{j=1}^{n}(\overline{y_{i.}} - \overline{y_{..}})(y_{ij} - \overline{y_{i.}})$$

$$\sum_{j=1}^{n}(y_{ij} - \bar{y}_{i.}) = y_{i.} - n\overline{y_{i.}} = y_{i.} - n\left(\frac{y_{i.}}{n}\right) = 0$$

0

$$\sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - \overline{y_{..}})^2 = n\sum_{i=1}^{a}(\overline{y_{i.}} - \overline{y_{..}})^2 + \sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{i.})^2 \quad \text{....} Eqn(1)$$

- This is the fundamental ANOVA identity
- It states that the total variability in the data as measured by the total corrected sum of squares can be partitioned into a sum of square of the differences between the treatment(factor level) average and the grand average, plus a sum of the squares of the differences of observation within treatments from the treatment average.

So, what we got is, again the same so whatever this expression, this expression is SS T. I am writing it again, this is my expression, you see, now this expression, I will do a bit of twisting I want something else. So, I will twist the expression a bit, but my expression should remain same, I should not change the expression in any way. But somehow I will do some sort of twisting. So, what I have done here, I have introduced one new term that is y i dot bar.

So, if I introduce one new term, then if I add one term, I will have to subtract the term as well, is not it? So that is what only I have done, I have added this and I have subtracted this also. So, by doing that, I have simplified it in this form earliest y ij - y double dot bar that is the grand mean. Now, I have introduced y i dot double bar what is y i dot bar remember, that is the mean of for one treatment one level mean of one level have introduced that new data mean of one level.

So, this I am adding it and subtracting so this is nothing but now it is in the form of $(a + b)^2$. So, $(a + b)^2$ is $a^2 + 2ab + b^2$. So that is what I got it. So, first, this is the term this term a^2 this is a, this is b, this is a^2. So, first I got this is my a^2. Say why I have written n for in this term, you will see there is no j there is only variable is only i so when I am writing j = 1 to n.

So, since there is no variable for j that means what it will be summation of those terms j times is not it? The n j = 1 to n time so that that is why I have written n into this from this I got n into this from this again here, y i and j both the terms are used. So, I am keeping both the summation. So, I got this term that is b^2 now 2ab so this is 2ab this term. Now see, now in this 2ab you say 2ab this term. So, I have taken a portion of this term portion of this 2ab which I have given here by red colour, you see I bought this term here, what do I get?

If so what is ∑ j = 1 y ij ∑ j = 1 y ij gives but they basically total of one level is not it? Total of ith level so it is y i dot and similarly, ∑ j = 1 y i dot bar somebody here no j term is there. So, basically it will be n x y i dot bar again what is y i dot bar? y i dot bar = y i dot divided by n. So, n n get cancelled, so, this term becomes 0 that means my this whole term becomes 0.

So, what is remaining is only $a^2 + b^2$ term. So, this is the equation to finally, I got this equation this is what my total variability of the data my overall variability of the data which is also called total character variance. So, this is equals to this is my a^2 term this is my b^2 term I got 2 components here and this equation let me call it as equation 1. Now, this equation 1 this is the fundamental ANOVA identity.

Now, we have seen what is ANOVA? ANOVA is something to be conscious of 2 variance one is variance between the group another is variance within the group another is based on that we have seen that but exactly in what way it is linked nothing we did not see now, we are coming to the ANOVA identity. So, this is the fundamental ANOVA identity. What does this state this expression if you look it by yourself, you do not have to look into the size just look at the expression a will be able to say.

What does it mean see what this is the total variability this portion? This is also red, my pen is also red, it is becoming difficult to this thing. So, now this portion is the total variability in the data. This is the total variability data this total variability of data is measured by these 2 what is this? This measured by the total corrected sum of squares can be partitioned into a sum of squares of the difference between the treatment average.

Difference sum of squares the difference between the treatment average seen in the difference between the treatment average what is the treatment average is? y i dot bar is the treatment average is not it? And y double dot bar is the grand mean. So, what is that is the sum of squares of the treatment average is not it? Basically it is the between group variability we are trying to find out the variability of treatment average from the grand mean exactly not variability.

Because when we find out variable it has to be divided by n - 1 term is the n - 1 term is missing just the sum of squares, this is the sum of squares of the difference between the treatment average and the grand average. There is a sum of squares between the treatment average and the grand average this is one term and what is the second term? This is the second term, what is the second term is a sum of squares of the differences of observation within treatment from the treatment average.

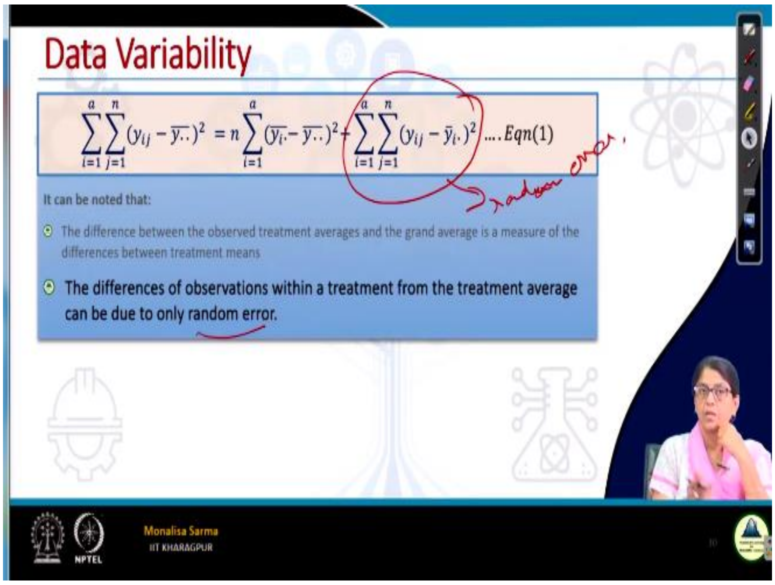
Say this is the treatment average from the treatment average what we are substituting each observation of the treatment. So, there are 3 treatment levels we will get 3 treatment means let me call it y 1 bar, y 2 bar, y 3 bar. I will get 3 treatment mean y 1 bar, y 2 bar, y 3 bar. So, now from the first treatment each value from each value I am substituting from y 1 bar what I am getting sum of all the plus again sum of each value from the sample 2 - y 2 bar.

That is a variance within that group, plus again variance within the third group within so that is my; this equation, this expression. So, understood what we got in this fundamental ANOVA identity. This total variability in the data which is measured as the total corrected sum of squares this I am working on the total variability in the data, is not it? I am let us not talk in terms of variability. Let us talk in terms of sum of squares.

So, what is this the sum of squares total collected sum of squares, because I am trying to find out the difference of each and every data from the mean, that is in my left hand side, in my left hand side, I am trying to find out the difference of each and every data from the mean. So that is my total corrected sum of squares. And this total corrected sum of squares I am partitioning it into 2 parts, what are the 2 parts what I did not partition this equation, I got the equations what is this 2 parts?

One part is that it is the sum of squares of difference between the treatment average and the grand average that is one part. That means here basically, I am finding out the difference between the between different groups and other is the sum of squares between each data within a group from the group mean.

(Refer Slide Time: 30:41)



Data Variability

$$\sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij}-\overline{y_{..}})^2 = n\sum_{i=1}^{a}(\overline{y_{i.}}-\overline{y_{..}})^2 + \underbrace{\sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij}-\bar{y}_{i.})^2}_{\text{random error}} \text{....} Eqn(1)$$

It can be noted that:

- The difference between the observed treatment averages and the grand average is a measure of the differences between treatment means
- The differences of observations within a treatment from the treatment average can be due to only random error.

So, this expression here again, I have written it the same expression. So, now, this expression, this difference between the observed treatment average and the grand average is a measure of the difference between treatment means, this portion, it is called a measure of the difference between treatment means, each mean and if each treatment means I am subtracting from the grand mean. Then, this is the difference of observation within a treatment from the treatment average this portion.

Within a treatment from the treatment average and this portion, why we get the variance within a sample why do we get a variance we have seen while doing the example for this concrete mixer, when mixing the concrete mixer, we are mixing with different chemicals to see the absorption ratio absorption amount basically. So, we have seen why there may be difference within the samples difference within the sample maybe the while doing the test, the environment was not the same for throughout the time that may be one of the reason.

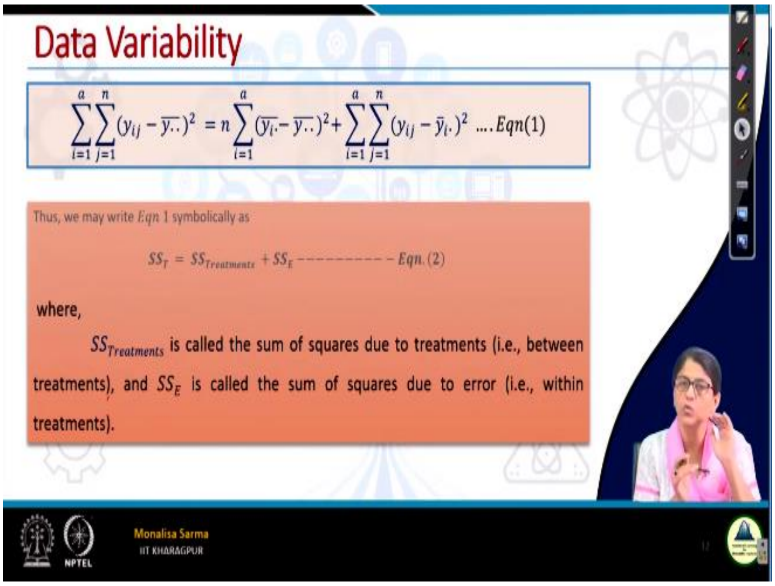
Or there may be certain the heterogeneity in the material itself or there may be a while the testing there may be testing equipments, we have used different testing equipment to test the whole lot of samples that different reasons. So, this is basically all these reasons we call it due to random error. So, this within the treatment from the treatment this whole thing, we can call it as random error is not it? Because variance within a treatment, it can be due to that only is not it?

Due to any of the reasons, and when was not same products for experiment heterogeneity of the material, heterogeneity of the measuring material, anything in any of the example if you take any other example things may be different, like when we are talking of different types of treatment for a particular disease, suppose for treatment a, treatment a group of people who are getting treatment a there also we might get variance among the data the curing time for some people may have cured getting to the 30 days.

Some people are getting cured 25 some 26 some 27 some 29 or maybe some 32, 35. So, they are maybe variance among themselves on the same treatment why many reasons maybe one of the reasons maybe the some people maybe immunity level is quite low some people immunity level is quite high and maybe think some people may not be taking the medicine properly very weak strong many other many other reasons like so, all those we can call it as a random error.

So, total avoidable variability in the data we are classifying it into 2 partitions one is maybe due to the random error and another is due to the treatment difference due to the treatment.

(Refer Slide Time: 33:55)



Data Variability

$$\sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij}-\overline{y_{..}})^2 = n\sum_{i=1}^{a}(\overline{y_{i.}}-\overline{y_{..}})^2+\sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij}-\bar{y}_{i.})^2 \text{ ....} Eqn(1)$$

Thus, we may write *Eqn* 1 symbolically as

$$SS_T = SS_{Treatments} + SS_E \text{ ----------- } Eqn.(2)$$

where,

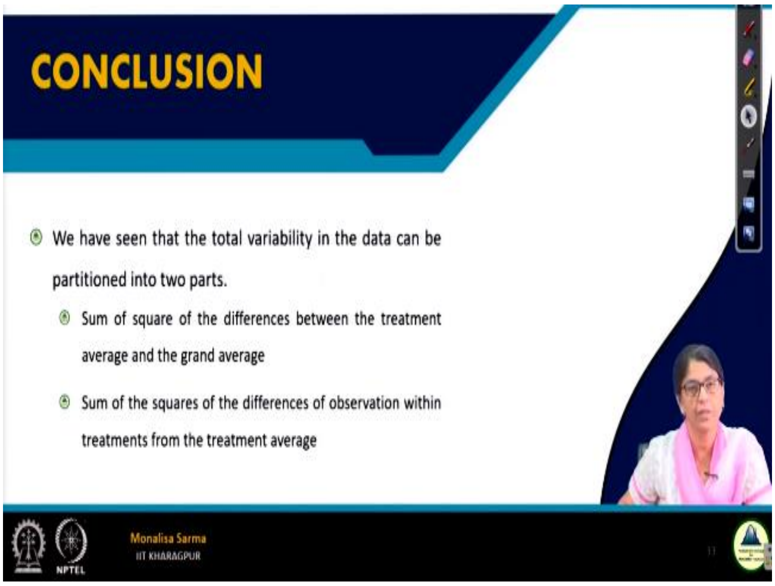
$SS_{Treatments}$ is called the sum of squares due to treatments (i.e., between treatments), and SS_E is called the sum of squares due to error (i.e., within treatments).

Monalisa Sarma
IIT KHARAGPUR

So, equation 1 we can write it symbolically as in this pattern. So, this is SS E, E in the subscript basically. So, SS T is the overall variable in the data SS treatment SS T meaning total variability in the data and SS treatment means variability among the treatment SS E is variable to due to random error. So, SS treatment is called a sum of squares due to treatment that is between treatment and SS E is called the sum of squares due to error that is within treatments, this is not the end of one way ANOVA.

So, we have till now we have come to this stage, where we found that the total variability in the data is partitioned into 2 parts. One part is the sum of squares due to treatment. Another part is the sum of squares due to errors. Still, there are many other things to discuss for this class. I am stopping my lecture here.

(Refer Slide Time: 35:00)



CONCLUSION

- We have seen that the total variability in the data can be partitioned into two parts.
  - Sum of square of the differences between the treatment average and the grand average
  - Sum of the squares of the differences of observation within treatments from the treatment average

Monalisa Sarma
IIT KHARAGPUR

So, see here if I want to conclude this, what I will say is that we have seen that the total variability in the data can be partitioned into 2 parts. Sum of the squares are the difference between the treatment average and the grand average sum of the squares of the difference of observation within the treatment from the treatment average.

(Refer Slide Time: 35:19)

REFERENCES

- Design and Analysis of Experiments (8th Edition), Douglas C. Montgomery, John Wiley & Sons, 2013
- Statistics, Witte, R.S., Witte, J.S., 2013., Wiley
- Probability and Statistics, 2nd edition, Moris DeGroot, 1986 Reading, MA: Addison-Wesley
- Freund, R.J. & Wilson, W.J. Statistical Methods, Academic Press, San Diego, 1997

Monalisa Sarma
IIT KHARAGPUR

So, these are the references than thank you guys.