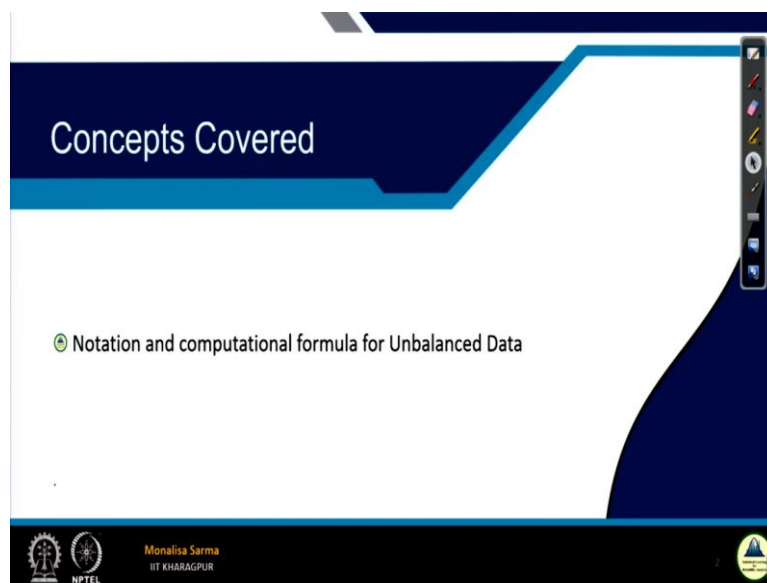


Statistical Learning for Reliability Analysis
Prof. Monalisa Sarma
Subir Chowdhury School of Quality and Reliability
Indian Institute of Technology, Kharagpur

Lecture - 37
ANOVA-VI

Hi everyone, so, last class we have seen some examples on ANOVA like how we can solve for ANOVA how we can find out whether the means of the different populations are same or they are not same with a particular significance level that is what we have seen.

(Refer Slide Time: 00:45)



And like till last class what we have discussed is that, if you have noticed all the samples that we have considered all the samples are of same size remember all the samples we have taken as n so basically, so, all the samples were of same size some today, we will just take a slight variance of that. So, it is not only possible that always we take a sample of same size, but it is advisable for ANOVA it is really advisable if it is better if we take samples of the same size.

But sometimes it is not possible we have to take samples of different sizes. So, today we will see the computational formula, if we use the samples of different sizes basically, we call that is unbalanced data.

(Refer Slide Time: 01:32)

Notation and Computational Formula

In essence, given a population a single factor of a levels, we have to calculate two estimations for σ^2 .

Sampling variance between groups with $(a - 1)$ degree of freedom	Sampling variance within groups with $(N - a)$ degree of freedom
$MS_{Treatment} = \frac{n \sum (\bar{y}_i - \bar{y})^2}{(a-1)}$	$MS_E = \frac{\sum_{i=1}^a [\sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2]}{\sum_{i=1}^a (n_i - 1)}$ $= \frac{\sum_{i=1}^a SS_i}{\sum n_i - a}$

Monalisa Sarma
IIT KHARAGPUR

So, this is what we have already seen in essence given a population a single factor of a levels, we have to calculate 2 estimate of σ^2 sampling variance between groups with $a - 1$ degree of freedom that is from that we get his MS Treatment. So, then sampling variants within group with $N - a$ degree of freedom that is we call it MS E.

(Refer Slide Time: 01:57)

Notation and Computational Formula for Unbalanced Data

In some single-factor experiments, the number of observations taken within each treatment may be different.	Number of samples (or levels) = a	
Such design is called unbalanced.	Number of observations in i^{th} level = $n_i, i = 1, 2, \dots, a$	
In such cases, slight modifications must be made in the sum of squares formulas.	Total number of observations = $N = \sum_i n_i$	a, n
	j^{th} observation in i^{th} sample = $y_{ij}, j = 1, 2, \dots, n_i$	y_i, y_j
	Sum of n_i observations in i^{th} sample = $T_i = \sum_{j=1}^{n_i} y_{ij}$	
	Sum of all n observations = $T = \sum_{i=1}^a T_i = \sum_{i=1}^a \sum_{j=1}^{n_i} y_{ij}$	

Monalisa Sarma
IIT KHARAGPUR

Then the F is MS treatment / MSE. So, now in some single factor experiment, the number of observation taken within each treatment may be different in some cases, it may not be same due to many other factors of course, our target should always be tried to keep this sample size same, because there are many what to say first now, you may ask why we should keep the same equal sample size first of all see when you have assumed that in remember when we tried solving for 1 way ANOVA we have taken 1 assumption.

We have taken 3 assumption basically out of them 1 assumption is that all the populations variance, we have considered it to be equal that is σ^2 is not it? All the population variance we have considered it to be equal that is σ^2 . And for that σ^2 , we found out a pooled variance estimate. So, now, if we it is sometimes it is really not possible that all the population variance may be exactly the same.

It is not exactly same, but on when we say same means it is similar and the same range maybe, but sometimes it is that it is not possible to be in a similar range. Sometimes it is a bit from the range. So, if we take the size of the sample as same this slight deviation of variance of this population between different samples, this will not have a very bad effect means this will not make our results incorrect.

This will not make our results this will not make to lose or precision of our results. But if the sample size differs then what happens slight change of a variation of these different populations, it may reflect too much on the our end results. And which makes our result in precise, so, as I told you, it is really not possible at all the variance will be almost same, there will be slight changes. So, that is why it is always better if we take samples of the same size.

The slight variations, if the sample size is not same, it will be reflected in the results and basically we will not get a result which we can be reliability of result will not be high. The result will not be precise result there will be error to it, our type 1 error will increase and type 1 error increase means definitely that is not something which we deserve. That is the first thing. Second thing one more important thing is that if our data is balanced if our sample sizes, say n for all the samples.

Then what happens our power is also maximum that is one way it see if our type 1 error remains low definitely our power will also be more is not it? So when we for unbalanced data what happens our type 1 error increases accordingly our power also reduces. So, it is a direct connection basically anyway I will not show you the total derivation of how the power increases for a stable data and unstable data how a power goes down that I will not deduce and show you but you know how to calculate power we have seen it.

For 1 population we have seen it similar way it can be done actually and I suggest you not to try also those simplification because it is a long computational and there are many software's

to that, we will which will do that when software is available why unnecessarily try doing lots and lots of calculation to find that so, there are many software's by the help of software you can just feed this data and you can see for the thing for the same the significance level for balanced and unbalanced data the same as the power.

So, now, as I told you this is called unbalanced design in such cases slight modification must be made in this sum of the squares formula when that is cases is when the our design is unbalanced whatever formula for f we have there will be a slight modification in the formula that is obvious right because there in our formulas what we have used? We have directly used n now, here for all the different samples they will be different n it will be n_i basically.

So, you see here how is the change in a formula number of samples is a number of observation in the i th level is n_i , so, then total number of observation is again N here N is $\sum_{i=1}^a n_i$ that is number of levels. So, earlier for n is what is N ? N directly we could write a into n total number of levels into a number of what to say observation in each level a into n .

So, now n will be this so, j th absorption in i th sample that is same and with j will go from 1 to n_i and sum of n_i observation i th sample so, sum of n_i observation i th sample here, this will be j will go from 1 to n_i it will not j it will not from the i goes to 1 to n it will go from 1 to n_i for each level, there will be different and maybe same but different for unbalanced data this is the formula instead here I am writing it is at T_i actually, if you can remember sum of all the observations in the earlier formula what I was using?

I was using is y_{ij} remember so actually now the coming formula which I will be using this y_{ij} this dot dot things it becomes difficult to represent. So I will just use a different nomenclature that is all but it is the same thing, so I am using a different nomenclature that is T_{ij} . So do not get confused the same thing, this time using a different variable that is all. So T_{ij} is a $j = 1$ to n_i y_{ij} than the sum of all n observations remember it was $y_{..}$ I use $y_{..}$ in the earlier. So, instead of that I have used a simple T , T is that will be $\sum_{i=1}^a T_i$ this $i = 1$ to $j = 1$ to n_i y_{ij} .

(Refer Slide Time: 08:42)

Notation and Computational Formula

The computational formula

<p style="text-align: center;">Total sum of squares</p> $SS_T = \sum_i \sum_j y_{ij}^2 - \frac{T^2}{N}$	<p style="text-align: center;">Between samples sum of squares</p> $SS_{Treatment} = \sum_i \frac{T_i^2}{n_i} - \frac{T^2}{N}$	<p style="text-align: center;">Within samples sum of squares</p> $SS_E = SS_T - SS_{Treatment}$
---	---	---

Monalisa Sarma
IIT KHARAGPUR

So, this difference will be reflected in their computational formula also, so, how it will be reflected so, SS_T is now this, this you can do it by yourself and we will see in SS_T earlier what was in SS_T in that place now, here you see y_{ij} , y_{ij} will be j will go from $j = 1$ to n_i for each i there will be different n_i is not it? n_1, n_2, n_3 . So, there will be changed accordingly yourself will be able to solve a simple calculations so, you will get SS_T is this.

So, $SS_{Treatment}$ similarly $SS_{Treatment}$ is this I am not doing the derivation you will be able to do it by yourself. So, this is the $SS_{Treatment}$ then and once we know SS_T once we know $SS_{Treatment}$ again SS_E we do not need to separately find it out SS_E is nothing but $SS_T - SS_{Treatment}$ what is SS_T , SS_T is the overall sum of squares. So $SS_{Treatment}$ is the sum of squares between treatment a sum of squares within group I can find out from this formula $SS_T - SS_{Treatment}$, so once I know SS_E , once I know $SS_{treatment}$.

I can find out the $MS_{treatment}$ I can find out the MS_E , is not it? What will be MS_E ? MS_E will be $SS_E - \text{total samples } N \text{ minus what? } N - a$, and $SS_{Treatment}$ will be $N - 1$.

(Refer Slide Time: 10:07)

Notation and Computational Formula


A mean square (or unbiased variance estimate) is given by
(sum of squares) ÷ (degrees of freedom)


e.g. $\hat{\sigma}^2 = \frac{(x - \bar{x})^2}{N-1}$

Hence


Total mean square,	$MS_T = \frac{SS_T}{N-1}$ ✓
Between samples mean square,	$MS_{Treatment} = \frac{SS_{Treatment}}{a-1}$ ✓
Within samples mean square,	$MS_E = \frac{SS_E}{N-a}$ ✓

Note that for the degrees of freedom: $(a-1) + (N-a) = (N-1)$ ✓





Monalisa Sarma
IIT KHARAGPUR



So, that gives sum of square divided by the degrees of freedom gives us the variance estimate, is not it? So, total mean square is MS_T is this $SS_T / N - 1$ $MS_{Treatment}$ is $SS_{Treatment} / a - 1$ and MS_E is $SS_E / N - a$ these are same just that there will be a slight difference of this value and this SS_T and $SS_{Treatment}$ you can just you take the earlier formula and as earlier formula instead of $j = 1$ to n there, you will just replace the $j = 1$ to n i and simplify it you will get it.

So, degrees of freedom is what $N - 1$ is the degrees of freedom of what, which one, do you remember, $N - 1$ is the degrees of freedom of the total variability of the data total variability of data degrees of freedom is $N - 1$, $a - 1$ is the degrees of freedom of the between treatment variables and $N - a$ is the degrees of freedom within each treatment.

(Refer Slide Time: 11:29)


Example 1: Using Formula for Unbalanced Data


For the previous example on 60W electric light bulbs, using the computational formula for unbalanced data:

Brand			
	1	2	3
16	18	26	
15	22	31	
13	20	24	
21	16	30	
15	24	24	


➔

Brands	1	2	3	Total
n_i	5	5	5	$N = 15$
T_i	80	100	135	$T = 315$





Monalisa Sarma
IIT KHARAGPUR



So, now take an example first, we will see that same example the last class we have seen for that 60 Watt bulb there are 3 different brands. And then we went on to prove that the 3 brands are we have enough what to say sufficient proof that the 3 brands are not same is not it? So, now, though this is balanced, because the sample size is the same of course, when it is a balance or non-balance, we can use the formula for non-balance.

But for non-balance, you cannot use the formula for balance because there are n changes, is not it? So, now, we will use the same example of a 60 Watt electric bulb using we will do it the same problem using the unbalanced data we will see we will get the same results or not. So, here for on further from that we found out what is n i is same for that is 3 and we found out the value of T i, T i is nothing but the summation of all these values remember.

(Refer Slide Time: 12:29)

Example 1: Using Formula for Unbalanced Data

		Brand		
		1	2	3
Table 1		16	18	26
		15	22	31
		13	20	24
		21	16	30
		15	24	24

Brands	1	2	3	Total
n_i	5	5	5	$N = 15$
T_i	80	100	135	$T = 315$

a) $\sum \sum y_{ij}^2 =$ Sum of squares of all entries in Table 1

$= 7045$

$\therefore SS_T = \sum_i \sum_j y_{ij}^2 - \frac{T^2}{N} = 7045 - \frac{315^2}{15} = 7045 - 6615 = 430$

So, now what is this? $\sum y_{ij}^2$ this is the sum of all squares of all the entries in table 1 is not it? y_{ij}^2 then this will give us the value 7045 then what does SS_T ? SS_T is $y_{ij}^2 - T^2 / N$ what is T^2 ? T is 315 sum of all this T is the grand sum is not it? So 315 so, this is n is 15 we got the SS_T we got 430.


(Refer Slide Time: 13:04)



Example 1: Using Formula for Unbalanced Data

Brand		
1	2	3
16	18	26
15	22	31
13	20	24
21	16	30
15	24	24

Brands	1	2	3	Total
n_i	5	5	5	$N = 15$
T_i	80	100	135	$T = 315$

$$\begin{aligned}
 \text{b) } SS_{\text{Treatment}} &= \sum_i \frac{T_i^2}{n_i} - \frac{T^2}{N} \\
 &= \frac{80^2}{5} + \frac{100^2}{5} + \frac{135^2}{5} - \frac{315^2}{15} \\
 &= 6925 - 6615 = 310
 \end{aligned}$$




Monalisa Sarma
IIT KHARAGPUR


So, similarly, now we will find out SS treatment. SS Treatment this is the formula $T_i / n_i T^2 / N$. So we have we found out T_i what is T ? T is the grand total and T_i is the total of within each treatment is not it? Earlier for T_i we use y_i dot again I am repeating for T we use y dot dot. So, similarly we find found out using the putting the value here we found out the value of the SS Treatment once you know SS Treatment and we can find out MS Treatment SS Treatment / $a - 1$ that is number of treatment - 1.


(Refer Slide Time: 13:47)



Example 1: Using Formula for Unbalanced Data

Brand		
1	2	3
16	18	26
15	22	31
13	20	24
21	16	30
15	24	24

Brands	1	2	3	Total
n_i	5	5	5	$N = 15$
T_i	80	100	135	$T = 315$

$$\text{c) } MS_{\text{Treatment}} = \frac{SS_{\text{Treatment}}}{a-1}$$




Monalisa Sarma
IIT KHARAGPUR


So and that is MS treatment, so once we have we found.

(Refer Slide Time: 13:55)

Example 1: Using Formula for Unbalanced Data

Brand		
1	2	3
16	18	26
15	22	31
13	20	24
21	16	30
15	24	24

Brands	1	2	3	Total
n_i	5	5	5	$N = 15$
T_i	80	100	135	$T = 315$

d) $MS_E = \frac{SS_E}{N-a} = \frac{SS_T - SS_{Treatment}}{N-a}$

$$= \frac{430 - 310}{15 - 3}$$

$$= \frac{120}{12} = 10$$

Note that $F = \frac{MS_{Treatment}}{MS_E}$

$$= \frac{155}{10}$$

$$= 15.5$$

Now we need to find out MS E for our MS E we can find out SS T - SS Treatment and / the degrees of freedom N – a. This we found 10 and we calculated it and we found 15.5. If you can remember, this the same value we got when in our last class when we have solved for this same problem 15.5 now with the same degree of significance when the significance level at one person.

We will see that this portion this value false in the rejection region. So we got the same values. That means what the formula for unbalanced data we can very well use for the balance data as well. But the formula for balanced data we cannot use for unbalanced data.

(Refer Slide Time: 14:44)

Example 2

⦿ In a comparison of the cleaning action of four detergents, 20 pieces of white cloth were first soiled with India ink. The cloths were then washed under controlled conditions with 5 pieces washed by each of the detergents. Unfortunately three pieces of cloth were 'lost' in the course of the experiment. Whiteness readings, made on the 17 remaining pieces of cloth, are shown below.

Detergent			
A	B	C	D
77	74	73	76
81	66	78	85
61	58	57	77
76		69	64
69		63	

⦿ Assuming all whiteness readings to be normally distributed with common variance, test the hypothesis of no difference between the four brands as regards mean whiteness readings after washing.

Now we will see a different example. This is an unbalanced example. See in the comparison of the cleaning action of 4 detergent 20 pieces of white cloths were first soiled with India ink

some type of ink. The cloths were then washed out the control condition with 5 pieces wash by each of the detergents there are we have taken 4 different types of detergents and we have soaked total 20 pieces of cloth with some type of ink.

And for each we are trying to we will wash it with 4 different types of detergent to find out there which detergent is best basically. So, unfortunately what happened unfortunately 3 pieces of cloth will last in the course of the experiment, So, definitely our data becomes unstable here, we have tried to keep the stable data because we know ANOVA it is always better if you keep the same number of samples the sample size same.

So, whiteness readings made on the 17 remaining pieces of cloth are shown below. So, for detergent A we got 5 pieces, good for detergent B we got 3 pieces, C we got 5, D we got 4 very much an unbalanced data for detergent type of B we got 3, detergent type of D we got 4 as you mean all whiteness reading to be normally distributed with common variance is the hypothesis of no difference between the 4 brands as regard mean with whiteness reading after washing.

So, we have to test the hypothesis that the mean what to say whiteness readings, whiteness, how white is the cloth after washing is there suppose there is some scale whiteness scale based on that scale that mean reading as is the mean reading of all the detergents same or is it different, that is what we need to find out basically. So, our null hypothesis is mean of all this is same alternate hypothesis at least 2 are not same.

(Refer Slide Time: 16:40)

Solution

- $H_0: \mu_i = \mu$ all $i = 1, 2, 3$ ✓
- $H_1: \mu_i \neq \mu$ some $i = 1, 2, 3$ ✓
- Significance level, $\alpha = 0.05$ (say)
- Degrees of freedom, $v_1 = a - 1 = 3$,
and $v_2 = N - a = 17 - 4 = 13$ ✓
- Critical region is $F > 3.411$

The slide also features a graph of a normal distribution curve. The x-axis is labeled 'F' and has a critical value of 3.411 marked. The area to the right of 3.411 is shaded and labeled 'Critical region reject H_0 ' with a '5%' significance level. The area to the left is labeled 'Accept H_0 '. A small inset video shows a person in the bottom right corner.

Monalisa Sarma
IIT Kharagpur

So, now, this is the null hypothesis, this is the alternate hypothesis, this is the significance level consider a significance level of 5% that is 0.05 then degrees of freedom of course, v_1 is 3 number of treatment is 3 - 1 sorry, it will be 4 different brands of detergent 4 - 1 is 3 and v_2 also 17 - 4 is 13. So, we found a critical region with for 5% significance level and with degrees of freedom 3 and 13.

5% significance level with degrees of freedom 3 and 13 the value of F is 3.411 So, now, we will this is from the table we got this is the value this is the critical value of F. So, any value greater than that will fall in the critical region, rejection region so, now from this data, we will find out the what value of F we get? We will find out MS Treatment we will find out MS E, $MS_{Treatment} / MS_E$ will give me the F value.

(Refer Slide Time: 17:47)

Solution

	A	B	C	D	Total
n_i	5	3	5	4	17 = N
T_i	364	198	340	302	1204 = T

$$\sum_i \sum_j y_{ij}^2 = 86362$$

$$SS_T = 86362 - \frac{1204^2}{17} = 1090.47$$

$$SS_{Treatment} = \left(\frac{364^2}{5} + \frac{198^2}{3} + \frac{340^2}{5} + \frac{302^2}{4} \right) - \frac{1204^2}{17} = 216.67$$

$$SS_E = 1090.47 - 216.67 = 873.80$$

Monalisa Sarma
IIT KHARAGPUR

So, we found out first y_{ij}^2 we found out this as the n_i , n_i is the sample size for these 3 samples then we found out T_i sum of all the samples of sum of all the observation of first detergent some of the all the 3 observation of second detergent this is the T_i value and this total is the T some of all this is T so, for our SS_T is y_{ij}^2 and this $T^2 /$ degrees of freedom.

So, this is what our SS_T value is not it? Our degrees of freedom are for degrees of freedom SS_T what is the degrees of freedom? Overall variability in the data overall variability in the data what is the degrees of freedom? Degrees of freedom is N and it is whole total number of observations of all the brands together minus the number of total levels is not it? So here what is the total number of observations?

Total number of observation is 17, and then how many brands are there 4 brands are there is not it? So then we found out the SS Treatment we found out similarly we found out SS E. What is SS E? $SS T - SS \text{ Treatment}$ will give us SS E.

(Refer Slide Time: 19:11)

Solution

The ANOVA table is now as follows.

Source of variation	Sum of squares	Degrees of freedom	Mean square	F ratio
Between detergents	216.67	3	72.22	1.07
Within detergents	873.80	13	67.22	
Total	1090.47	16		

- The F ratio of 1.07 does not lie in the critical region.
- Thus there is no evidence, at the 5% significance level, to suggest a difference between the four brands as regards mean whiteness after washing.

Monalisa Sarma
IIT KHARAGPUR

So between the detergents we found this is the sum of squares this degrees of freedom is 3 to 4 different samples. Within detergents this variance we found is 873. This is the value Within detergents is this one, this one also this is between treatment. This is the within treatment 873, 873 within the treatment 873 than this degree of freedom is 3 this degrees of freedom is 13. Then we found the F value is 1.07, 1.07 is the F value.

And what is our critical reason, our critical reason is 3.411 then 1.001 is something here in this range. So, it very much lies in the acceptance ranges that means our null hypothesis is not rejected. So F ratio does not lie in the critical region does, there is no evidence at the 5% significance level to suggest a difference between the 4 brands as a result, mean whiteness after washing. So, there is no evidence to suggest that there is a difference, difference of whiteness in the different 4 brands.

(Refer Slide Time: 20:40)

CONCLUSION

- ☉ In this lecture we have seen analysis of variance calculation for unbalanced data. ✓
- ☉ For equal sample size the test statistic is relatively insensitive to small departures from the assumption of equal variances for the 'a' treatments, but this is not the case for the unbalanced data.

Monalisa Sarma
IIT KHARAGPUR

So that is what we have what we have seen for ANOVA I do not think now, I do not think initially I started with ANOVA discussion telling that it is a bit complicated, but now, I think you will not tell that this is complicated, it is very easy just find out the 2 variance estimate. One thing is completed this is to understand how to find out the 2 variance estimate. If you just try to do it; using the formula that will be very easy, but why do not; do that, why we are using this formula that you should understand.

That is what we have tried to explain in all the lectures why we are using the how this formula has come when we are trying to find out the variance within treatment, when you are trying to find out the variance what to say between treatment and why we are trying to find out this 2 estimate of σ squared and why we have used F distribution, why does it mean greater will give us a getting a value of F which is better than the critical region why does it is this what to say the null hypothesis is rejected all these things you should understand.

Just calculating it is very easy, and this is one thing this is what we have done this is for basically when you try to compare the means of the different population. Now, similarly, when you try to compare the variance of the different population, there is also another technique that is called MANOVA. So that we will not be discussing this is beyond the scope of this course. So, in this lecture what we have seen we have seen analyse variance and variance of pool unbalanced data.

And see as I have mentioned, for equal sample size, the test statistic is relatively insensitive to small departure from the assumption of equal variance. I have mentioned this is not it? For

equal sample size the test statistics whatever test statistics means here, the test statistic means that F value, when he call it a test statistics, it is when I am doing chi square distribution, my test statistic is the chi square value, when I am using F distribution, my test statistic is the F value t distribution my test statistic is the t value.

So here for equal sample size the test statistic is relatively insensitive to small departure from the assumption of equal variance for the a treatment for the a treatment, we have assumed that a variants are equal, but small departure from the assumption that it is equal it is if we take equal sample size, it is insensitive to it; it will not have what to say it will not have a results which will give us a which will inflate our type 1 error.

So, but this is not the case, but unbalanced data unbalanced data or type 1 error gets inflated if there is a departure from the assumption of equal variance. So, that makes the test if a slight variation if it is very sensitive to slight variance that means what? That means the test is not robust. If it is insensitive to small variance that means the test is robust. So, when our test is robust when we take the sample is of equal size.

So that is why it is always advisable to take for ANOVA takes sample size, equal sample size, but always it is not possible as we have seen in the example.

(Refer Slide Time: 23:54)



So these are the references and thank you guys. So we have completed an hour today. So we will be starting a new topic in my next lecture.