∑αβμσ√λ Statistical Learning for Reliability Analysis

Prof. Monalisa Sarma
Subir Chowdhury School of Quality and Reliability
Indian Institute of Technology - Kharagpur

Lecture - 36
ANOVA - V

Hello guys so, today to continue our discussion on analysis of variance, we have almost covered the required concept of ANOVA by that I mean only 1 way ANOVA because we have just discussed 1 way ANOVA the 2 way ANOVA 3 way ANOVA multi way ANOVA and all this I have not discussed as I told you, once you know that 1 way ANOVA you will be able to learn it by yourself only.

And moreover, it is really difficult to teach all those in a classroom and because intense computation is required for that, however, the concepts are the same with the same concept we will be able to do the rest.

(Refer Slide Time: 01:08)

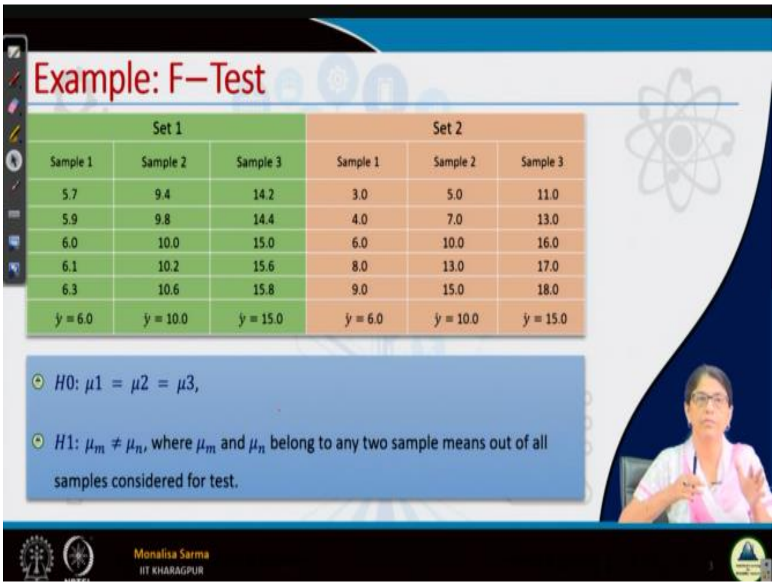
Concepts Covered

- Some examples on application of ANOVA

Monalisa Sarma
IIT KHARAGPUR

So, now in today's class, what we will be doing is that we will be seeing some examples, we have fine we have learned applications, we have learned the concepts. Now, we will see how we will use those to solve the problem.

(Refer Slide Time: 01:17)



Example: F–Test

Set 1			Set 2		
Sample 1	Sample 2	Sample 3	Sample 1	Sample 2	Sample 3
5.7	9.4	14.2	3.0	5.0	11.0
5.9	9.8	14.4	4.0	7.0	13.0
6.0	10.0	15.0	6.0	10.0	16.0
6.1	10.2	15.6	8.0	13.0	17.0
6.3	10.6	15.8	9.0	15.0	18.0
$\dot{y} = 6.0$	$\dot{y} = 10.0$	$\dot{y} = 15.0$	$\dot{y} = 6.0$	$\dot{y} = 10.0$	$\dot{y} = 15.0$

- $H0$: $\mu1 = \mu2 = \mu3$,
- $H1$: $\mu_m \neq \mu_n$, where μ_m and μ_n belong to any two sample means out of all samples considered for test.

The first problem that have examples, the first example that I have taken is it is not an that contrary data that data we which we have just created. That is it is not an example as such, we have remembered this data, we have used to come to the conclusion that is why when we want to compare means a different population, why do we need ANOVA? We have used this example for that purpose.

And now, so first example we will try this only with this example, we will see what ANOVA gives and what the result ANOVA gives in this contrary set of data. So, if you can remember properly there in the last class not I think the first class on ANOVA what we have discussed is that there are 2 sets of data that is set 1 and set 2. And we see there are 3 difference we have taken 3 samples from 3 different population maybe we have taken from 3 different population or maybe not.

We do not know like we have to prove that whether all the populations are really different. That is what it means that the populations are really different that is what we wanted to prove that was our objective. So since we see here in the set 1 sample 1 we got some mean of 6 first let us see what we got in sample 1 we got mean 6 sample 2 we got mean 10 sample 3 we got mean 15 this is quite a difference mean quite a significant difference in the all 3 sample.

Now, if we just tell from the mean, you have we can see that and we can tell that, the means are really different that means these 3 samples are taken for the from 3 different populations having 3 different means, if we just take see the mean from the means we can see that similarly, the set 2 also set 2 we see the same set 2 the mean of sample 1 is same as the set 1 mean of sample 1 set, then mean of sample 2 of set 2 is same as mean of sample 2 of set 1, similarly for set 3, also.

So that is, that is why we could also tell for even for set 2, also, all the 3 samples are from 3 different populations whose means are different. And in fact, if we can just from the means we can say maybe that for both set 1 and set 2, sample 1, of set 1 and sample 2, maybe it is from the same population, again sample 2 have both set 1 and set 2, maybe from the same population.

Similarly, for sample 3 for both set 1 and set 2, maybe from the same populations, means 3 different populations from one population, we have taken sample 1 for both set 1 and set 2, for another population, we have taken sample 2 for set 1 and set 2, likewise. So from just looking at the means we could come to that conclusion. So now, remember, we were not very much convinced with that.

Because this difference; in means maybe attributed to many other factors. That is what we have discussed earlier. So, what we have tried to do remember we have plotted a boxplot for this. So when we have plotted the box plot for the set of data, what we saw is that for this sample 1, I do not have the figure right now with me it is in the first slide of the ANOVA and it was lecture first lecture basically.

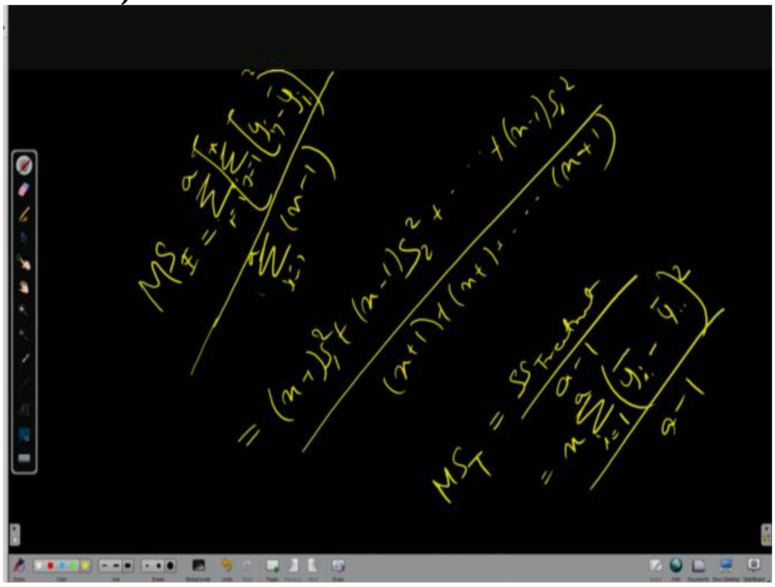
So what does the bar boxplot indicate boxplot indicate the 3 samples of set 1 there from the we could see they are very closely bunched together and that means the variance among themselves within variance, variance within the sample was very less and from the boxplot we could very easily say, this sample 1 sample 2 sample 3 these 3 may have come from different populations, but same we could not say for set 2.

Because it was it has been the huge variance it was as a from the figure as if we can say that as if it has come from a single population 3 different sample, but it has come from a single population it was its variance is so high for all the sample 1 sample 2 sample 3, that means, we could tell that for set 1 3 means are different, but for set 2, we could not say that a 3 means are different, why?

Because we know from sampling distribution of mean we know that if the variance is high, if we have a high variance, that the mean of the sampling distribution is equal to the population mean that statement is not a very precise statement reliability of that statement is not high, we cannot say with accurate precision that the sample mean is equal to the population mean when the variability is quite high variability indicates the standard error among different means.

Sampling distribution of mean means what? Sampling distribution of means, means that different means have different samples, isn't it? So, now, the same example now, we will use to do this, we will be using the same application to find out using ANOVA whether what we got from the boxplot is that the same result we get it or not, let us see that. So, first thing is that if we try doing the ANOVA first what we need to find out for that, you remember ANOVA thing.

(Refer Slide Time: 06:25)



$$MS_E = \frac{\sum_{i=1}^{a}\left[\sum_{j=1}^{n}(y_{ij}-\bar{y}_{i.})^2\right]}{\sum_{i=1}^{a}(n-1)}$$

$$= \frac{(n-1)S_1^2+(n-1)S_2^2+\cdots+(n-1)S_a^2}{(n-1)+(n-1)+\cdots+(n-1)}$$

$$MS_T = \frac{SS_{Treatment}}{a-1} = \frac{n\sum_{i=1}^{a}(\bar{y}_{i.}-\bar{y}_{..})^2}{a-1}$$

So, if you can remember what was our F value of F we need to find out the value of F. So, what was F if you remember correctly, f is MS treatment. Treatment by MS error, this MS treatment was the difference of treatment mean from the grand mean and MS error is for all those different for all the samples basically we tried to find out the variance within the sample so if you just a quick recap.

So what was the formula for MS error we call it MS E basically. So, what was MS E? MS E = ∑ i = 1 to a indicating that there are total a levels than summation of in each level how many data days was to 1 to n. So, (y ij - y i dot bar) ^{2} and divided by the i = 1 to a n - 1.

Remember this was the formula for MS E basically if you can this also you can write it, it was nothing but a pool variance estimate also distinct.

We can also write it in this way n - 1 s ^{2} + n - 1 s 1 ^{2} basically s 2 ^{2} + n - 1 s i ^{2} then n + 1 + n + 1 + n + 1 this is equals to this we have seen it isn't it? What is this n + 1 up to total how many levels 14 A levels and this as 1 ^{2} is the variance of the individual level. So, n is the number of data in each sample here the samples all the samples are of the same size that is n. So, this is saying we have already seen it.

So, this is how we can calculate MS E we can either calculate MS E by this formula or we can calculate MS E by this formula whatever is this the same thing from this we got this and similarly, what was MS T MS treatment that is MS treatment is equals to SS treatment / a - 1 that is the level minus 1 and this is what was that n ∑ i = 1 to a it is the sum of squares of this treatment means from the grand mean so, it is why i dot - y dot dot this is the grand mean y dot dot whole square / a - 1.

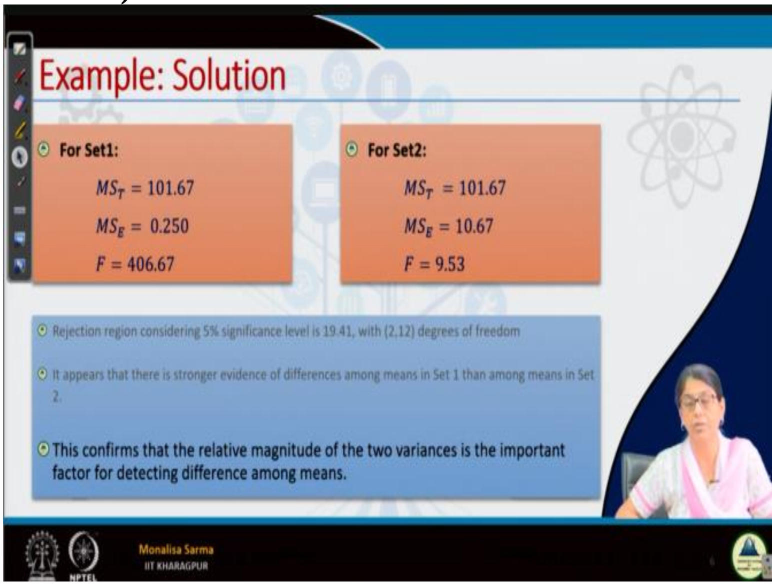
So, this was our formula remember, so, now, for this sample set, if we first thing is that we will calculate what is that MS treatment as well as MS error. So, if you want to calculate MS treatment, what do you need to do? I will just find out a variance from this from each sample I will find a variance So, n - 1 very this is suppose these variances as suppose this variance is s 1 ^{2} this variance of this is s 2 ^{2} variance of this is s 3 ^{2}.

So what is how many 1, 2, 3, 4, 5, 5 is n is 5 7 - 1. So, n - 1 s 1 ^{2} n - 1 s 2 ^{2} + n - 1 s 3 ^{2} divided by how much n - a n - 1 + n - 1 + n - 1 that is n - a n - 3. So, that way we will get MS error similarly, we can find out the MS treatment, MS treatment formula is this the treatment we will find out the treatment mean this is the treatment mean find out what is the grand mean of the grand mean is 6 + 10 + 15 / 3 that will be the grand mean, is not it?

So, from the each mean each treatment mean we will subtract a grand mean for all the levels then divided by the degrees of freedom. So, that way we will get the values of course, the first thing is that we need to know what is the hypothesis we will be able to solve it using the hypothesis testing only is not it the first thing is what is the hypothesis? So, for us the hypothesis is μ 1 = μ 2 = μ 3 that means.

All the populations come from means of all the 3 populations are same that is the null hypothesis. Alternate hypothesis is at least 2 of the means are not equal. So, for that we will have to find out the F ratio. So, if value of f falls in the critical region, then we reject the null hypothesis if the value of L f does not fall into critical reasons, then we cannot reject the null hypothesis.

(Refer Slide Time: 11:29)



Example: Solution

- For Set1:
  - $MS_T = 101.67$
  - $MS_E = 0.250$
  - $F = 406.67$
- For Set2:
  - $MS_T = 101.67$
  - $MS_E = 10.67$
  - $F = 9.53$

- Rejection region considering 5% significance level is 19.41, with (2,12) degrees of freedom
- It appears that there is stronger evidence of differences among means in Set 1 than among means in Set 2.
- This confirms that the relative magnitude of the two variances is the important factor for detecting difference among means.

So, by doing this calculations, we got MS T is equal to this MS E is equal to this and we got the value of F is this for this this is for the set 1 set 1 we got this value. Similarly, we will do for set 2 also set 2 we got this values F value you see F = 9.53 now, directly from seeing the F value only you can make it out the difference F will remember F distribution is something of this sort and more the value of F means it is going this way more the value of F means F will be this way, and F is 9.53 means f is very much in this range, isn't it?

And as I told you for ANOVA remember it was always a 1 tailed distribution we are only interested in finding out if it is greater than we are not interested in finding out less than. So, because if all the means are not same, that means our MS T value will be quite big compared to the MS E value. So if MS T value will be big that means for more MS T value is more the bigger more the value of F.

So, we are not interested in finding out the lower tail isn't it so lower till when we will get when MS T value is less than only we will get the lower tail and MS T value will be lesser than the MS T value that is very unlikely we are no we are not interested in finding an MS E

MST value will be bigger when the treatment means are different isn't it? So, we are only interested in the upper tail.

So here what is the rejection region if we consider at 5% significance level? So at 5% significance level what is the degrees of freedom here degrees of freedom is for MS T what is the degrees of freedom MS T degrees of freedom is a - 1 that is total level is 3 so degrees of freedom is 2 and for MS E degrees of freedom is n - a and it is 15 total value is 15, 15 - 3 is 12. So, our degrees of freedom is 2 - 12.

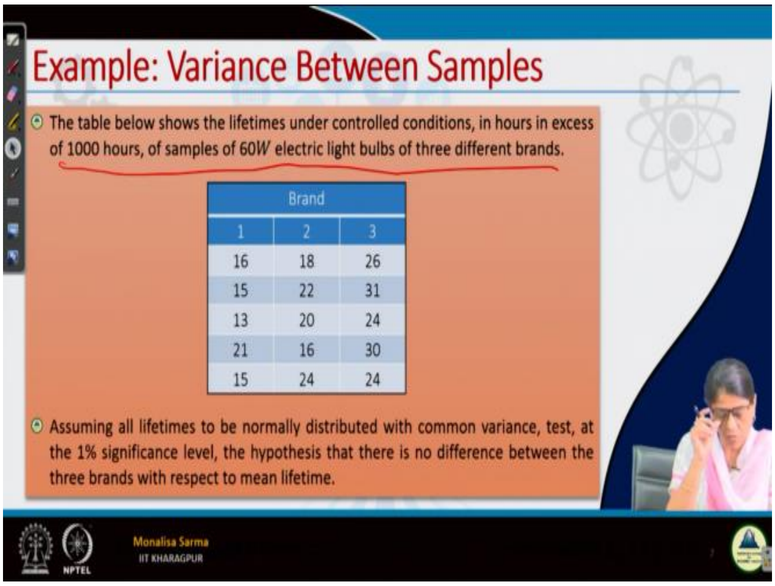
So, if you see the F table for 5% significance level see here we will consider α = 2.05 only we will not take α = 2 α / 2 will not do α / 2 because it is single tail it is not 2 tail. So, if we see the table for that we will get the this is develop a test 19.41, 19.41 is the minimum if it is greater than 19.41 then it falls under critical region and see what is the value of F in for the first set F is 406.67. It is much, much greater than 19.4.

And it is very much in the critical region. Even a boxplot also sought from a boxer also we could see that and for the set 2 we got in the acceptance regions 9.53. So though the means of both the sets are same, but still for set 1 we could very conveniently say that all the means are different for the 3 different population null hypothesis is rejected. And for set 2 null hypothesis, we could not reject the null hypothesis.

We do not have enough evidence to reject the null hypothesis. Clear that is how we do the ANOVA testing. So, it appears that there is a stronger evidence of difference among means in set 1 then among means instead to stronger evidence. Since it is very much in the critical region, this confirms the relative magnitude of the 2 variants is the important factor for detecting differences among means.

We have already seen that isn't it, that is why only we have learned ANOVA this was the why we need to do in our just for our justification, we have seen this example.

(Refer Slide Time: 15:20)



Example: Variance Between Samples

- The table below shows the lifetimes under controlled conditions, in hours in excess of 1000 hours, of samples of 60*W* electric light bulbs of three different brands.

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

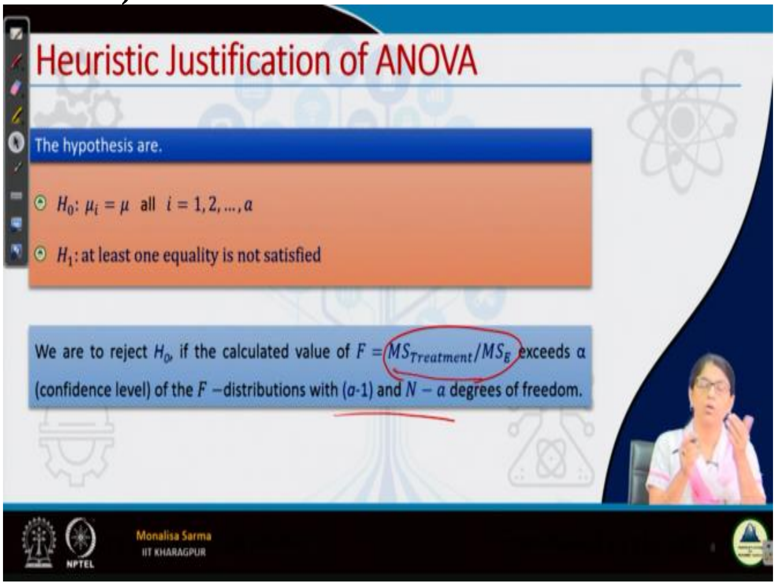
- Assuming all lifetimes to be normally distributed with common variance, test, at the 1% significance level, the hypothesis that there is no difference between the three brands with respect to mean lifetime.

Monalisa Sarma
IIT KHARAGPUR

Now, so, let us take another example, here that table below it shows the lifetime under control condition in hours in excess of 1000 hours of a sample of 60 Watt electric light bulbs of 3 different brands, 3 different brands of lights there it is. So, it gives us the hours and life of the bulb life or the bulb in excess of 1000 hours. So, we need to find out whether the 3 brands are same on the means of 3 brands a lifetime of all the 3 brands are same or not.

So null hypothesis is μ 1 = μ 2 = μ 3 alternate hypothesis is at least 1 of μ 1 μ 2 μ 3 at least 2 are not same so that is the alternate hypothesis. So, we have to test that one person significance level that means 0.01.

(Refer Slide Time: 16:17)



Heuristic Justification of ANOVA

The hypothesis are.

- H_0: $\mu_i = \mu$ all $i = 1, 2, \ldots, a$
- H_1: at least one equality is not satisfied

We are to reject H_0, if the calculated value of $F = MS_{Treatment}/MS_E$ exceeds α (confidence level) of the F –distributions with (a-1) and $N - a$ degrees of freedom.

Monalisa Sarma
IIT KHARAGPUR

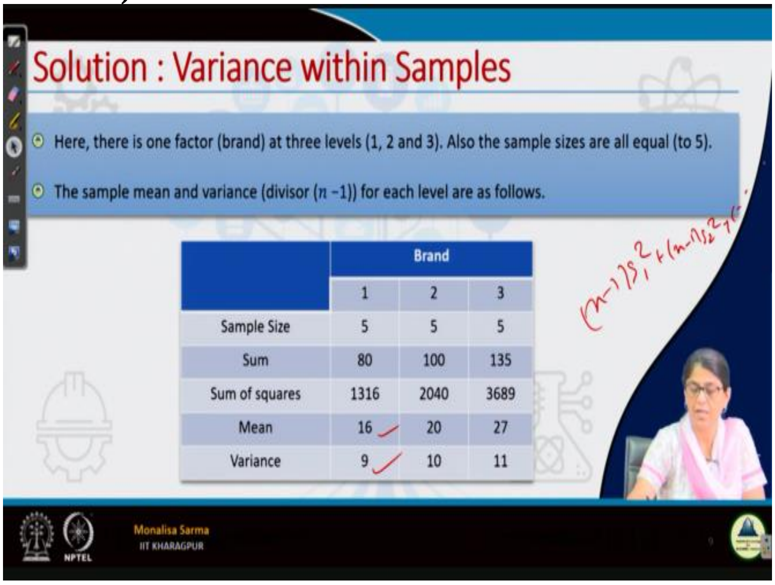
So what are the hypothesis here? Hypothesis is μ is when i for all i = 1 to a. How many here a is equals to just 3. So an alternate hypothesis is at least 1 equality is not satisfied. So we have to reject H 0 if the calculated value of F exceeds A confidence level of the app distribution with a - 1 and n - a degrees of freedom, isn't it MS treatment degree of freedom is a - 1 MS E degree of freedom is n - a.

Remember when we have learned F distribution that time it is a general convention in the table it is given that the table follows all the standard textbooks in the table it follows that for the greater the numerator, which is bigger, we keep it in the value of the variance which is bigger we keep it in a numerator that was a while trying to compare population variance of 2 different populations.

But for when we use app distribution for ANOVA, then we do not have to find out which one is bigger and which one will keep in numerator always the numerator will be MS treatment. Because if there is because it does not a difference in the means definitely MS treatment will be more if there is no difference in the mean MS statement will be almost similar to MS E but there of course there will be a slight difference but it will be almost similar to MS E.

So we do not have to bother whether we will have to put the bigger one a smaller one always MS treatment will be MS treatment will always be bigger or may be similar to MS E so an MS treatment is always put in a numerator.

(Refer Slide Time: 18:05)



Solution : Variance within Samples

- Here, there is one factor (brand) at three levels (1, 2 and 3). Also the sample sizes are all equal (to 5).
- The sample mean and variance (divisor $(n-1)$) for each level are as follows.

	Brand		
	1	2	3
Sample Size	5	5	5
Sum	80	100	135
Sum of squares	1316	2040	3689
Mean	16 ✓	20	27
Variance	9 ✓	10	11

$(n-1)s_1^2 + (n-1)s_2^2 + (\ldots$

So now so there is only 1 factor what is the factor that different brands and 3 levels 1, 2, 3 and a sample size is equal to 5 first we will find out the sample this is a good way this actually does I am why have solved with this way any given problem whenever you try to solve it this way, just draw make a table because then it becomes very easier to calculate make a table first.

And then different levels, put the different levels then sample size for each of course, now we are discussing the sample sizes same for all the samples and of all different populations then find the sum of all the samples values, sum of all the sample values means this well the sum of all these values, sum of all these values, sum of all these values. So, this is the sum then sum of squares, sum of squares means what is the sum of squares.

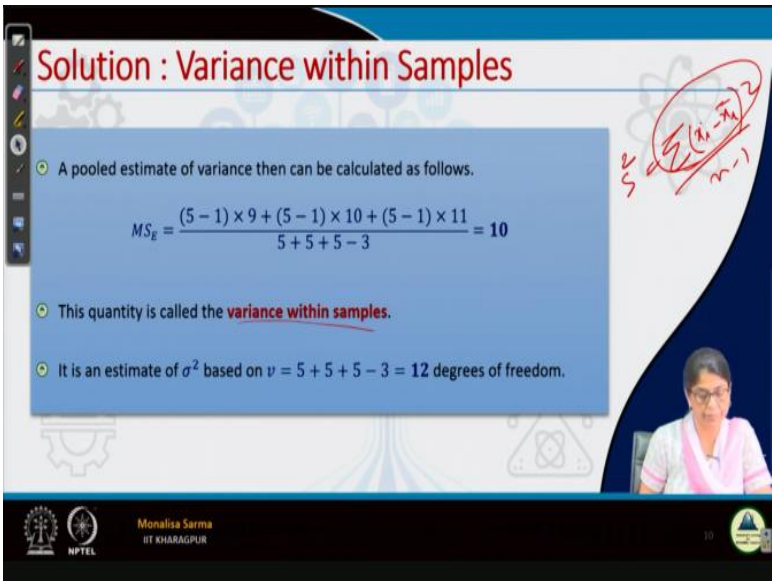
Actually here I did not mention it very properly sum of squares, there are some squares of different types one is the when we are trying to find out the sum of sum of squares between within the variance, isn't it? I mean within one sample, that how do we find out the sum of squares of that, that is, we take each data of a sample subtracted from the mean of the sample mean of the treatment isn't it?

Each data of the treatment substituted from the mean of the treatment and square that is the sum of squares for each treatment. And that is within groups sum of squares basically, and one is between groups sum of squares, how do we find that between groups sum of squares. So, basically, it is better if you have 2 column here, one is within group sum of squares and next is between groups and squares and between groups sum of squares.

How do we find out that is we will subtract them treatment mean from the grand mean and add up all those will suppose we have 3 treatments. Then treatment mean of the sample 1 - grand mean + treatment mean of the sample 2 - grand mean + treatment mean of the sample 3 - grand mean, of course square. So that is sum of squares between treatment. So instead of 1 column here, which I shown as 1 column.

And so it is always better if you do have a 2 column sum of squares between groups and sum of squares between groups then is the mean then you find a variance. Now since once we found a variance, finding out the MS E will be very easy for us, isn't it? What does MS E MS E n - 1 s 1 ^{2} + n - 1 s 2 ^{2} n - 1 s 3 ^{2} how many and divided by n - a. So it will be very easy for us to calculate MS E.

(Refer Slide Time: 20:49)



Solution : Variance within Samples

- A pooled estimate of variance then can be calculated as follows.

$$MS_E = \frac{(5-1)\times 9 + (5-1)\times 10 + (5-1)\times 11}{5+5+5-3} = 10$$

- This quantity is called the variance within samples.
- It is an estimate of σ^2 based on $\upsilon = 5+5+5-3 = 12$ degrees of freedom.

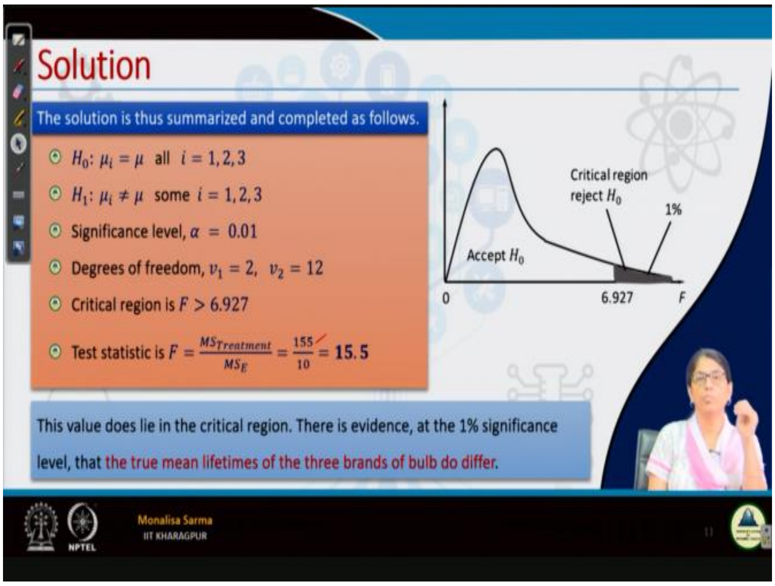
$$s^2 = \frac{\sum (x_i - \bar{x}_i)^2}{n-1}$$

So that is also called pooled variance estimate if you remember. So let us now say it was so easy when we calculate the first we calculate the variance also, if you try to do that, first it at first glance, it may seem it is very complicated, but it is no just find the variance of each sample. Once you find the variance, first find the mean of the sample variance of the sample. And then once you have the mean of the sample, then you find the grand mean.

Grand mean nothing but the mean of all divided by the total number of means, then finding out the sum of squares also becomes very easy sum of squares within sample when you are calculating variance you already calculate that, is not it? What is variance, variance is equals to a sample variance s ^{2} =, x i - x i sorry x i bar x bar basically whole square / n - 1. So this is variance. So when you calculate variance already, you calculate the sum of squares.

So this is MSE. So, what is MS E? MS E is the variance within samples then as we have already discussed, this is an estimate of is basically the pooled variance estimate, isn't it? Pooled variance estimate, and what is the degrees of freedom of the degrees of freedom is n - a n - number of levels, n is the or sorry, not n, n is the sample size capital and N means sum of all the sample sizes, sample sizes. So, B is degrees of freedom is here total number of item is 15 - 3 is 12.

(Refer Slide Time: 22:32)



Solution

The solution is thus summarized and completed as follows.

- $H_0: \mu_i = \mu$ all $i = 1,2,3$
- $H_1: \mu_i \neq \mu$ some $i = 1,2,3$
- Significance level, $\alpha = 0.01$
- Degrees of freedom, $v_1 = 2, \; v_2 = 12$
- Critical region is $F > 6.927$
- Test statistic is $F = \frac{MS_{Treatment}}{MS_E} = \frac{155}{10} = \mathbf{15.5}$

This value does lie in the critical region. There is evidence, at the 1% significance level, that the true mean lifetimes of the three brands of bulb do differ.

Monalisa Sarma
IIT KHARAGPUR

So, now, similarly, you find MS treatment also MS treatment, what will be the degrees of freedom for MS statement, the degrees of freedom for a misstatement will be just the level minus one number of level minus one a - 1. So find out the value of MS treatment. So here MS treatment, I did not calculate it and so you can do it very easily so it is 155. So, what is F Statistics we got 15.5.

And what is the significance level is 1% significance level. So, at the 1% significance level from the table, we are using the degrees of freedom 2 and 12. What is 2 is the degrees of freedom of MS statement and 12 is the degrees of freedom of MS E, using this degrees of freedom, if you see the F table, you will see that this value corresponding to this is 6.927. So, any value which is greater than 6.927 it falls in the rejection region or the critical region lesser than 6.927 means is the acceptance region.

So, what we got we got our F Statistics values 15.5. So, it is 15.5 means it falls in the rejection region. So, that means it is false in the rejection region means that means we have a we reject the null hypothesis that means all the 3 brands once we reject the null hypothesis definitely we accept the alternate hypothesis. So, that means the 3 brands or population means the 3 brands are different.

There is evidence at 1% significance level that the true mean lifetime so, the 3 brands of bulb do differ. So, this is a ANOVA, just understanding the ANOVA a bit difficult actually it is not difficult if you try to understand it properly. If you please go once if you cannot understand it once please go through the slides 2, 3 times learn from slide number 1 and

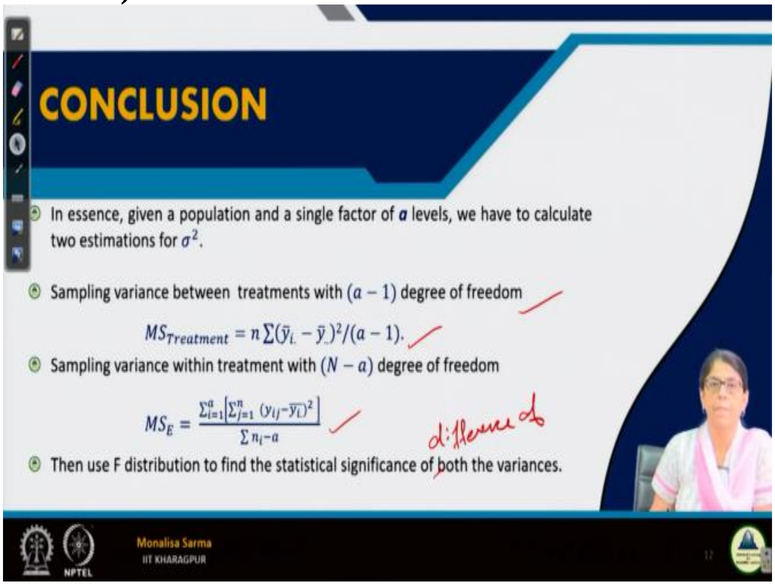
ANOVA I think I have covered in till total 5 lectures and today's the 6th lecture on antibody here we are doing problems despite lecture go through it again and again.

And it will be very easy for you to calculate it is nothing first you should understand how what does actually ANOVA does first you need to understand that basically ANOVA gives us ANOVA entity that ANOVA entity gives us the 2 estimate of the variance. So 1 estimate of the variance we get it from the difference of treatment means, another estimate of the variance we get it from within a sample what is the variance this 2 variants and these are the 2 variance estimate.

So, if the population means are equal then what happened both the variance estimates are same and if the population is not it may not be exactly the same it is similar value and if the population variance or if the population means are not same, then we will get a value which gives them MS treatment basically, that means the variance within that treatment that is a quite a big value and which leads us to F value which falls in the critical region.

And if it falls in a critical reason, that means, we reject the null hypothesis that is null hypothesis is that the all the population means are not same. So, we have enough evidence to say that the all the population means are not see.

(Refer Slide Time: 25:55)



CONCLUSION

- In essence, given a population and a single factor of ***a*** levels, we have to calculate two estimations for σ^2.
- Sampling variance between treatments with $(a-1)$ degree of freedom

$$MS_{Treatment} = n\sum(\bar{y}_{i.} - \bar{y}_{..})^2/(a-1).$$

- Sampling variance within treatment with $(N-a)$ degree of freedom

$$MS_E = \frac{\sum_{i=1}^{a}\left[\sum_{j=1}^{n}(y_{ij}-\bar{y}_{i.})^2\right]}{\sum n_i - a}$$

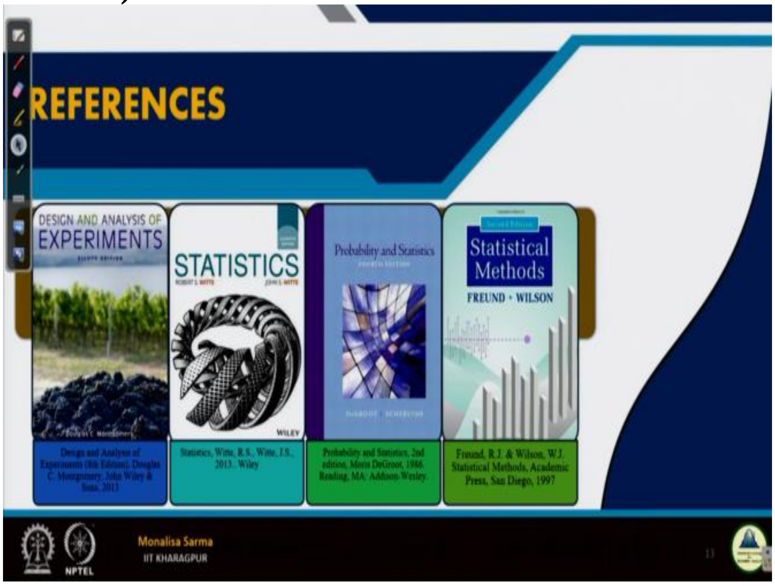
- Then use F distribution to find the statistical significance of both the variances.

difference of

Monalisa Sarma
IIT KHARAGPUR

So, now, this to conclude this in essence given a population and a single factor level factor of a levels, we have to calculate 2 estimates of σ ^{2} sampling variance between treatment with a - 1 degrees of freedom is MS treatment sampling variants within treatment with N - a degrees of freedom is this then use F distribution to find a statistical significance of both variants statistical difference are both variants basically. So, and in fact, this would be so, difference of both the variances.

(Refer Slide Time: 26:48)



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

DESIGN AND ANALYSIS OF EXPERIMENTS
STATISTICS
WILEY
Probability and Statistics
Statistical Methods
FREUND • WILSON
NPTEL

So, these are the references mainly I am covering from design and analysis of experiments. Other books also you can also refer to other books and thank you guys. Thank you.