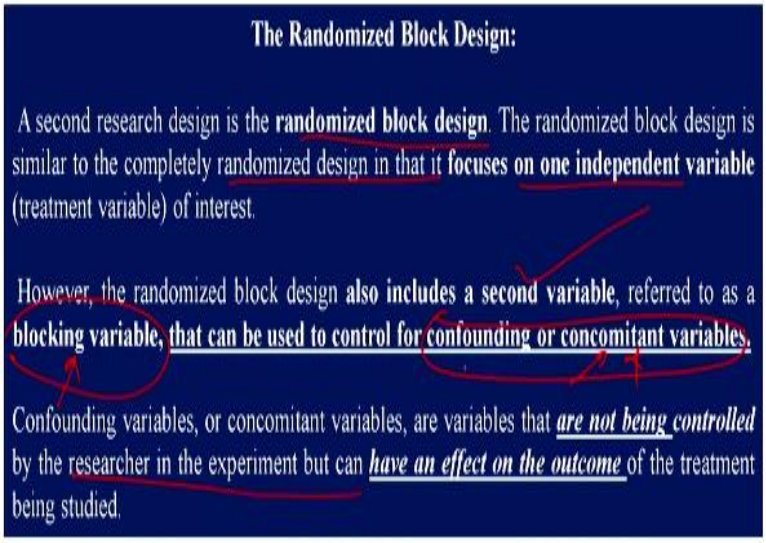
Business Statistics
Prof. M. K. Barua
Department of Management Studies
Indian Institute of Technology - Roorkee

Lecture – 50
Randomised Block Design

Hello friends, I welcome you all in this session, as you are aware in previous session we were discussing about the assumptions of ANOVA and one of the assumptions was the equality of variances of population from where we take samples and for that we applied Levene's test and we have seen that in Levene's test, the null hypothesis was that all variances were equal and alternative was, they were not equal.

So, we did not reject the hypothesis it means we concluded that the all variances were same and once that hypothesis is not rejected, when null hypothesis is not rejected, we are now in a position to solve a question on ANOVA, right so in today's class, we are going to talk about the second type of experimental research design, it is called randomised block design, as you are aware the first type of design was completely randomised design where we study the effect of treatments on dependent variable.

(Refer Slide Time: 01:59)

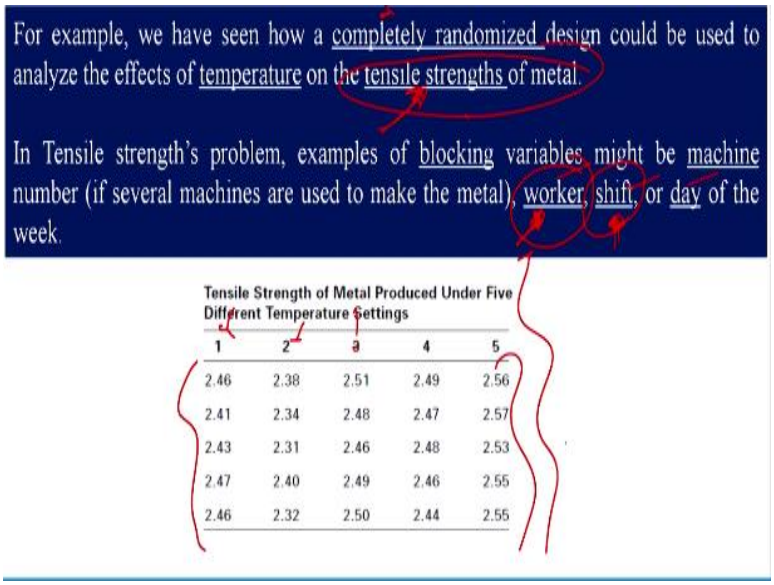


So, what we do in randomised block design is it is very similar to completely randomised block design which in fact focuses on one independent variable but here in randomised block design it

includes one more variable and this variable is referred as blocking variable, so this blocking variable can be used for controlling the confounding or concomitant variables, now what are these confounding or concomitant variables?

So, these are the variables which affect the output as we know that the dependent variable gets affected by independent variable but there are some variables which affect the dependent variable as well apart from independent variable, so these variables are actually not been controlled by the researcher in the experiment, so we need to control them in one way or the other.

(Refer Slide Time: 03:30)



So, blocking variable helps us in controlling confounding variables, we will take examples of these variables; confounding variables, so we have seen an example wherein we just tested the effect of temperature on tensile strength of metal using one way ANOVA and that was a case where we used completely randomised design, right. Now, if you look at this example then you will find that the tensile strength depends on several other factors; it is not only on temperature.

For example, let us say you have got different suppliers of raw materials, you have got different shifts, you have got different operators on the machine, you have got different weather conditions, you have got different humidity levels and so on, so there are several confounding

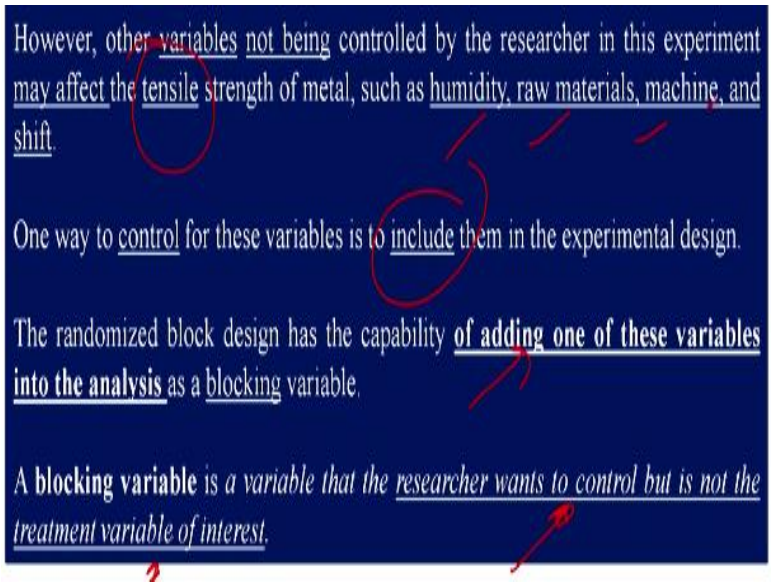
variables which affect the tensile strength which we have not controlled in the first part of experimental design which was completely randomised design, right.

So, you can have several such confounding variables which would be affecting dependent variable apart from independent variable, right, these are some of the examples of blocking or the confounding variables and we need to block all these suppose, if we block let us say worker, when I say worker means, the skill set worker possesses, so if we take worker and we block this skill set, so we will have workers of different skill set and temperature; different temperatures and then we will measure tensile strength.

Or we can check the tensile strength of metal, when the metal is being or let us say the operation is being done in different shifts and at different temperature levels, so this would be come in that case a blocking variable right, so this is the question what we have seen where in there were 5 different temperature conditions and this was the tensile strength, so we checked this whether the tensile strength has got or whether the temperature has got any effect on tensile strength.

So, this is what we have seen in one way ANOVA, so generally what we do in any experiment, we do not control confounding variables.

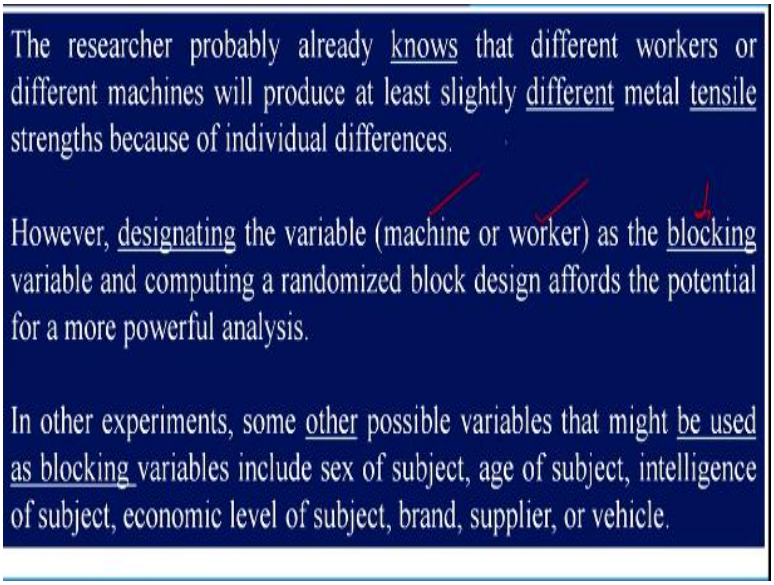
(Refer Slide Time: 06:28)



But actually they affect the dependent variable in this example, tensile strength and these are some of the confounding variables, right, so one way to control these variables is to include them in the experiment, so the randomised block design has the capability of adding one of these variables into the analysis as blocking variables, so we can use randomised block design wherein we would be using one of these confounding variables as a blocking variable.

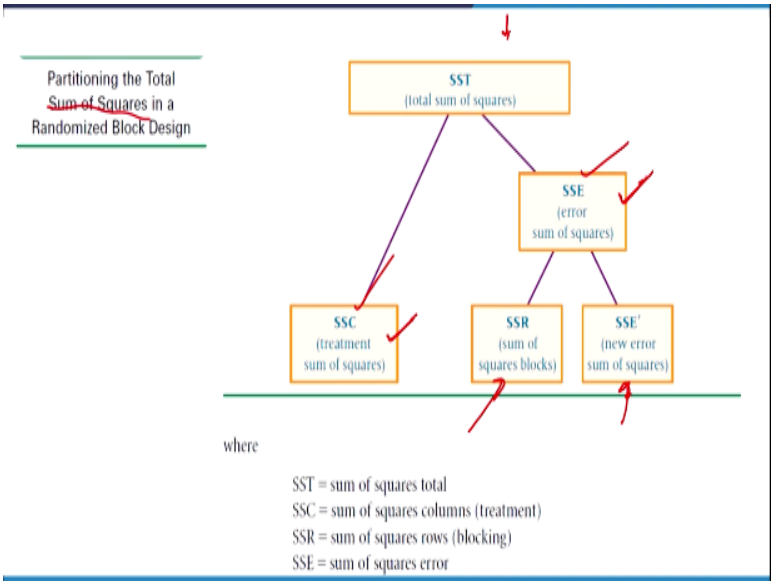
So, what is blocking variable; is a variable that the researcher wants to control but is not the treatment variable of interest, okay, so though we will focus only on treatment independent variable but we will have a blocking variable as well so, in reality what happens whenever you perform any experiment you know that the independent variable might affect dependent variable and there would be several confounding variables.

(Refer Slide Time: 07:53)



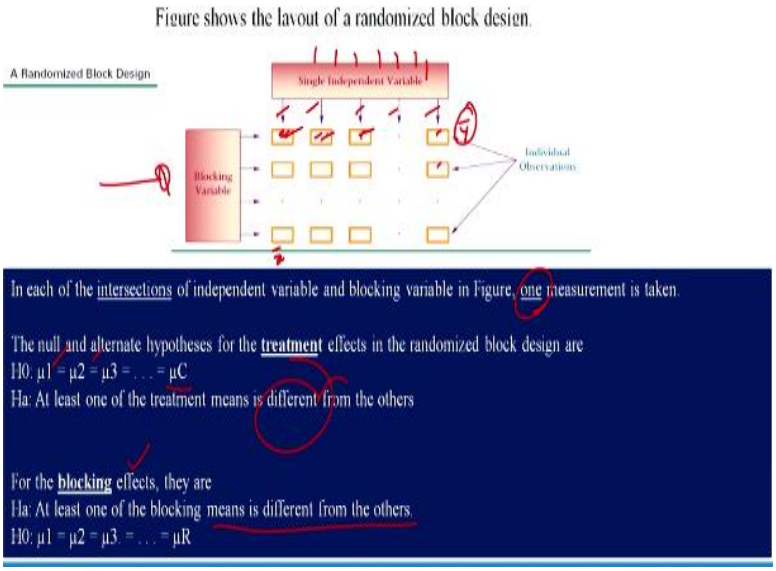
So, you will try to control those confounding variables, so as I said blocking is one of the methods of controlling confounding variables, so you can have any confounding variable let us say machine or worker or shift or let us say humidity or any other confounding variable, so that the confounding variable would be designated as blocking variable. Now, in n other experiments some other possible variables that might be used as blocking variables such as again, the gender of the worker, the educational qualification of the worker, the age of the worker, intelligence of the subject.

(Refer Slide Time: 08:50)



So, there can be n number of confounding variables, so this is how the partitioning of total sum of squares happens, so this is column sum of a square and this is error sum of a square which is again summation of this row sum of a square and this is error sum of squares, right so, if you calculate these 2, then you can easily find out sum of total squares.

(Refer Slide Time: 09:28)

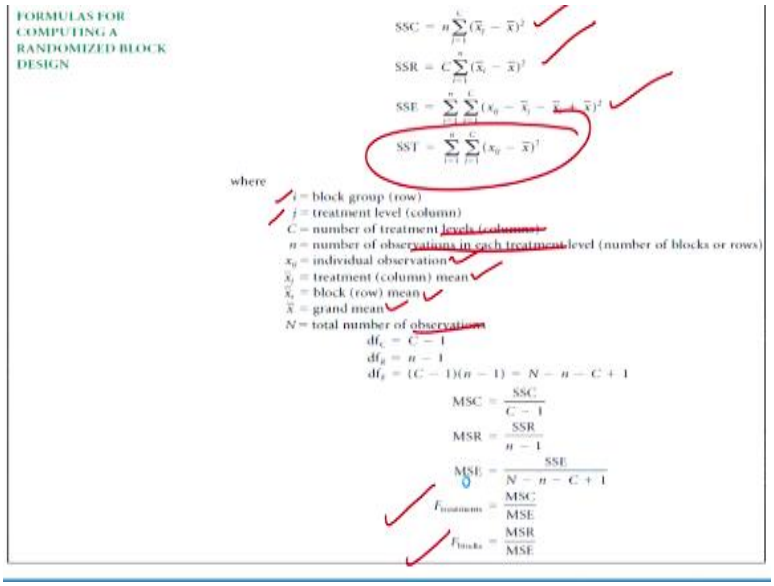


This how the layout of the randomised block design looks like, so you have got single independent variable, so over here, just one independent variable with different treatment levels and there is one blocking variable, right and these are different cells wherein you would be having different observations and that would be nothing but dependent variable, so here at each of these intersections, we are taking just one measurement.

So, you need to frame null and alternative hypothesis, so null hypothesis is that let the means of these columns are same, right, so μ1, μ2 and μ3 and all other means are same and alternative is they are not same, they are different from each other similarly, for blocking affect at least one of the blocking means is different from others, it means this mean, right, mean of these values.

So, you have got column mean over here, let us say x bar, here you have got let us say call it y bar for simplicity, right, so mean of row, so we are saying that all these means of rows are same and alternative is that they are not same, right but again our main focus is on the effect of independent variables on dependent variable, our main focus is not on whether these block means are same or this row means are same.

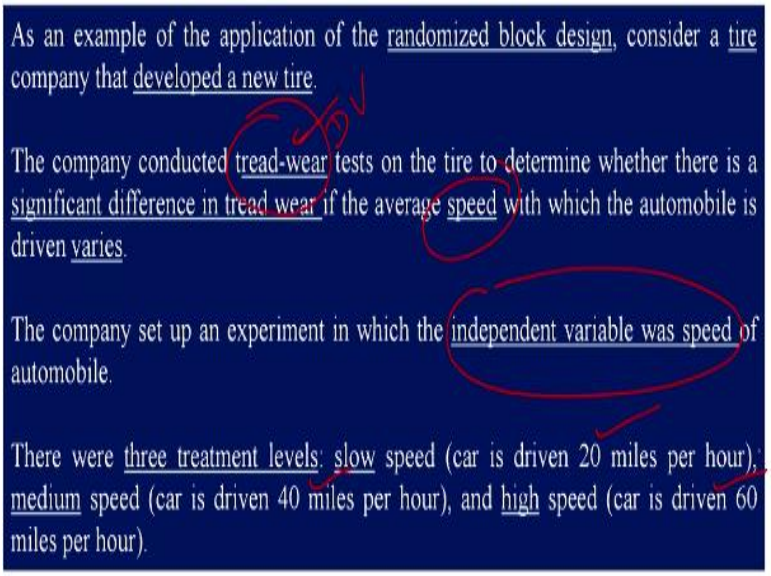
(Refer Slide Time: 11:29)



Main focus is on whether these column means are same or not, right, so this how you can calculate column sum of square row, sum of the square error, sum of a square and then total sum of squares, so you can have different let us say treatment levels, you can have block groups number of treatment levels, seen n is number of observations in each treatment level, these are individual observations, treatment mean, block mean, grand and total number of observations.

So, you need to calculate something called F value, so F for you will be calculating 2F values for treatment as well as for blocks, right.

(Refer Slide Time: 12:05)



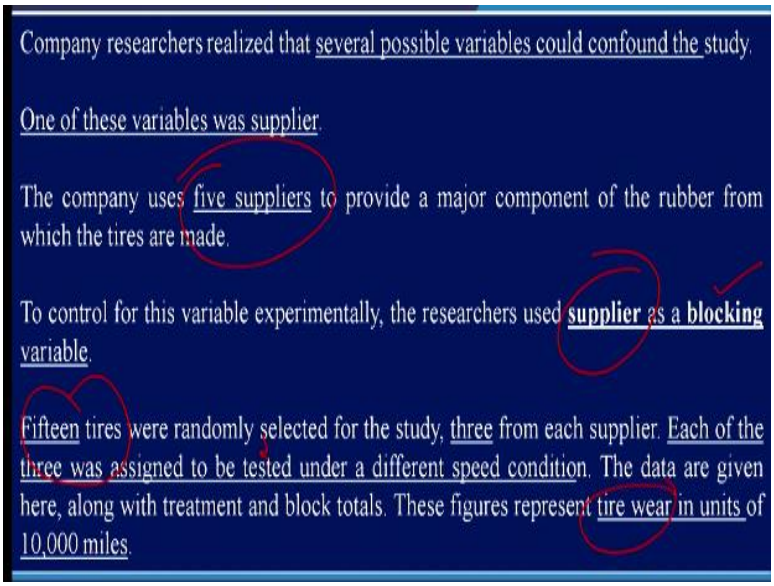
So, let us take an example of randomised block design, so let us take an example where there is a company which is developed new tire and the company wants to test or want to check tread wear of the on these tires, how these tread wears occur on tires with a varying speed of the vehicle, right so the company has set up an experiment where independent variable is speed and the dependent variable is nothing but this tread wear, right.

So, how speed is affecting tread wear, so this is your dependent variable DV, so there were 3 treatment levels, so in fact an experiment was set up and the 3 speeds were chosen, 20 miles per hour, 40 miles per hour and 60 miles per hour, so these are 3 different treatment levels of independent variable which is the speed, right. Now, we know that the tread wear also depends on several other factors not only on speed.

For example, let us say who are the suppliers of raw material for which tire was made, whom made the tires, in which shift the tires were made, what was the temperature conditions when those tires were made, what was the humidity condition, what was the let us say, educational level of the workers and so on so, in this experiment we have decided to block just one variable and that was raw material supply, right.

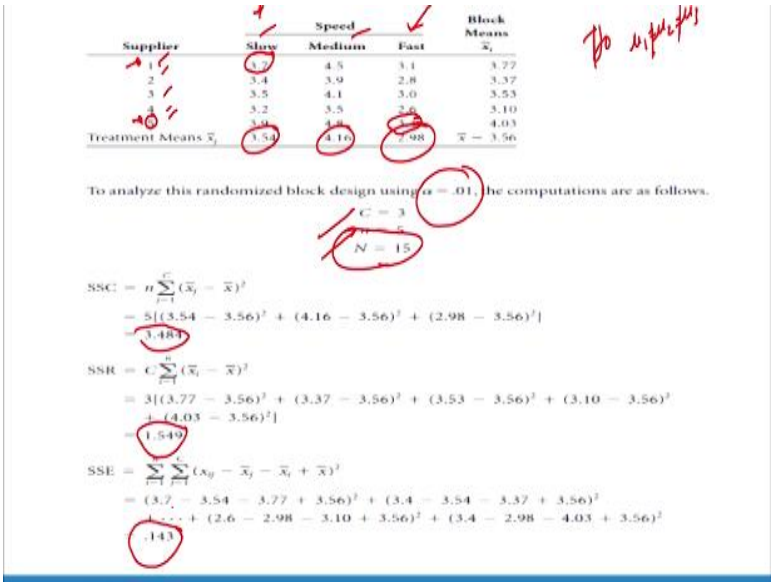
So, in this experiment there were 5 raw material suppliers and the company manufactured tire from those raw materials which they received from 5 suppliers.

(Refer Slide Time: 14:41)



So, we will say that the blocking variable is supplier, so what we have done in this experiment; what the experimenters have done is that they took 3 tires from each supplier, so total 15 tires were randomly selected for the study, so 3 tires from each of the suppliers, so total 15 tires, each of the three was assigned to be tested under different speed conditions, so the data on tire wear in terms of units of 10,000 miles are given in next slide.

(Refer Slide Time: 15:20)

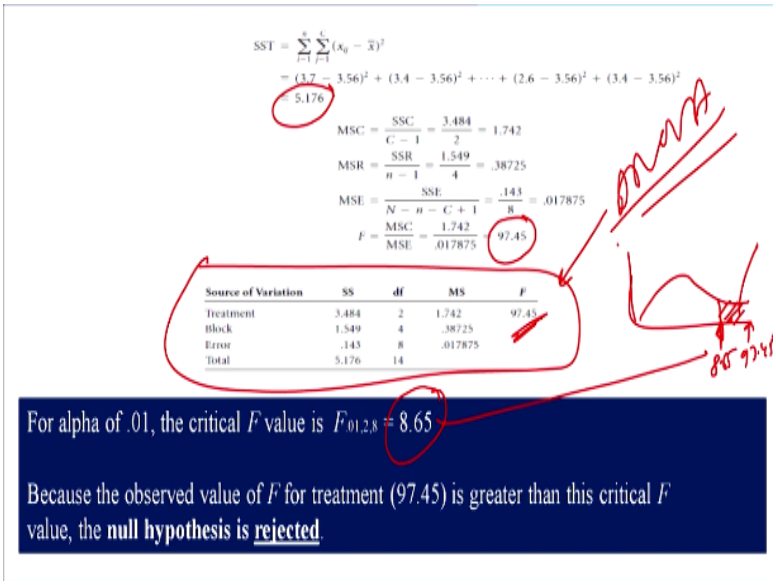


So, let us say there are 5 suppliers and 3 speeds, right. Slow, medium and fast speed, so this is tread wear when experiment was performed on tire manufactured from raw material supplied by supplier one and the speed was slow. What is this; the tread wear, the measurement are the dependent variable, tread wear, when the speed was fast and the tire was made from raw material supplied by supplier number 5.

So, we have to find out is there any significant difference in these means, so that is the question, right, so this is; so you can have treatment, a null hypothesis for treatment, tight that all let us say μ1, μ2 and μ3, all these column means are same, right and they are not same is alternative hypothesis, right, so you need to test this hypothesis at 99% significance level or alpha = .01 right.

So, we know that there are 3 groups, these 3 levels, there are 5 suppliers and total 15 observations, so you need to calculate column sum of squares, row sum of squares and error sum of squares.

(Refer Slide Time: 17:09)

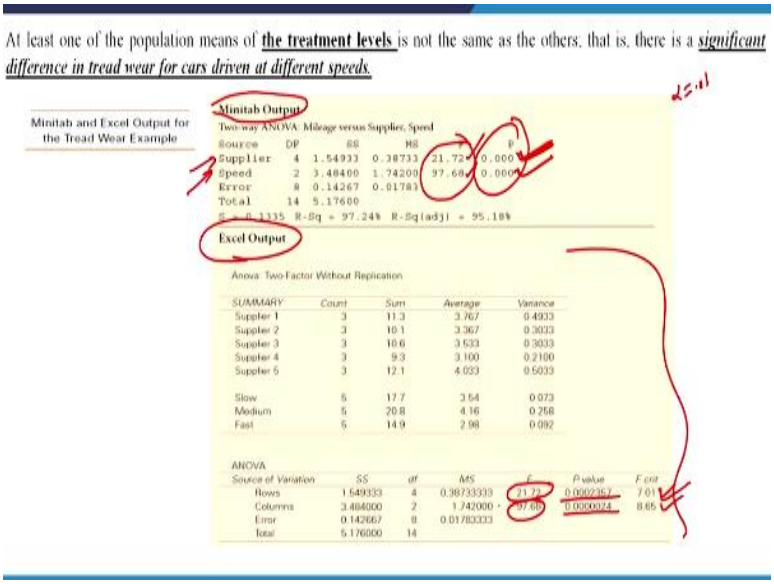


So, if you know all these 3, we can calculate total sum of a square, right which is this, 5.176, right, now let us find out what is the value of F, which is 97.4, this is output from; this is ANOVA table, right, ANOVA output table, so F is 97.45, it is much, much higher than the value

1, right, so your distribution would look like this, right, so this is your rejection region, right and this value is let us say 97.45 and this is your critical value here actually, right.

So, this critical value is in fact is to be found from table, right, it is 8.65, so we will reject null hypothesis, it means that at these 3 different steps; tread wear or different, mean number of tread wear are different right, so we have rejected null hypothesis, right.

(Refer Slide Time: 18:24)

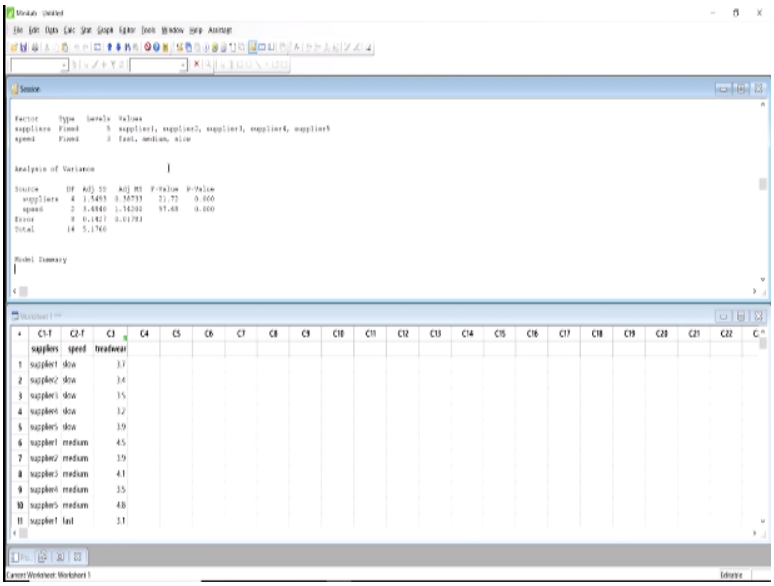


So, there is a significant difference in tread wear for cars driven in different speed that is our conclusion and this is the Minitab output where p values are < alpha, right, .01 was the alpha, so here both these p values, so we will say that there is a significant difference in means of tread wear of different suppliers right, as well as different speed, we will go back to our question, so the mean; this means they are also significantly different.

That is why you are getting 2p values or 2 F values in ANOVA table, right, just see this 2F values, okay, so we will say that there is a significant difference in tread wear as far as speeds are concerned in as far as the suppliers are concerned, right, so this is the output from Excel, right, same answer just see this p values for let us say supplier is 0, this is supplier means row, right, so we will just see, see this is as good as 0.

Then for a speed which is column, right, this is for column, right, we will use again approximately 0, F values these are critical values of F, right, so which are, these are critical values and these are actual values, right, so you can see that these F values are greater than critical value, so you need to reject null hypothesis, okay.

(Refer Slide Time: 20:20)



So, we will work out this question using Minitab and let us see whether we get the same answer or not, so you have got suppliers and speed and this is tread wear, right, so there are 5 suppliers, right, so we can write supplier 1, supplier 2, supplier 3, supplier 4 and supplier 5 and if you look at in fact this is to be repeated several times because there are total 15 observations, so 15, let us look at speed.

So, the first is slow right, then medium and then, okay you can have in fact 5 slow, 5 medium and 5 fast, right and then next fast, right, so now you need to enter tread wear values, right so supplier 1 speed slow, what was the tread wear; 3.7, supplier 2 speed slow 3.4, supplier 3 speed slow 3.5, similarly you are supposed to write for supplier 5, when speed is slow, now after entering all these points, we need to enter data for all the suppliers when speed is medium, right.

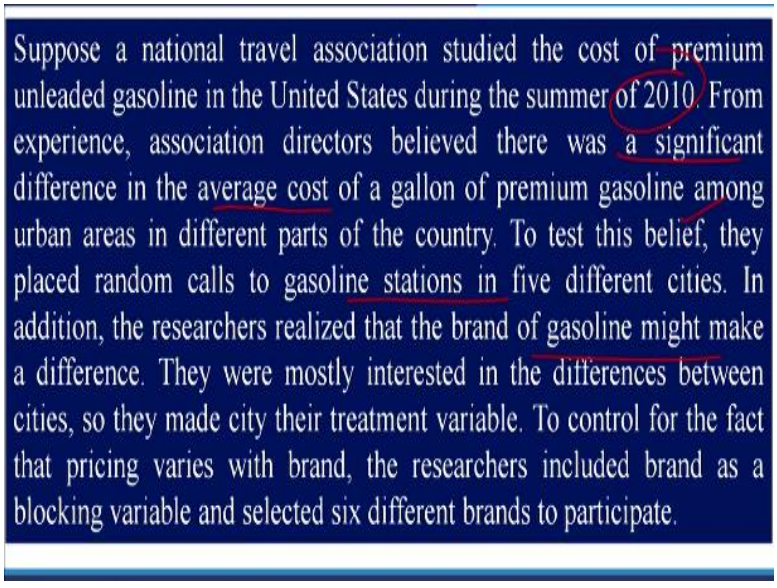
And after that you need to enter data for all 5 suppliers, when speed is fast, so you have got this data entry done now, we will go to stat ANOVA, go to generalise general linear model, this is the point you need to remember, right, general linear model fit general linear model, right, so you

have got response is nothing but your dependent variable right and factors are not your independent variables, right.

So, these are factors, right, so we will go to let us say options to check and the significance level, so this is to be tested at I think what percentage, at what significance level, it was 99% significance level right, okay, we will get the answer now, so this is the output, we will look at the F this ANOVA table first, just look at this, for supplier it is 0, p values for speed also 0, right, so we will reject null hypothesis.

Because these two are p values is < alpha for supplier as well as for speed, right and these are F values, now if you look at; if you go downward, just let us see what are other things there in the output, okay in fact, the other things are not of use, so you just look at the F values and p values, right, so you are rejecting null hypothesis, okay, so this how you should solve a question, we will move on to 2nd question.

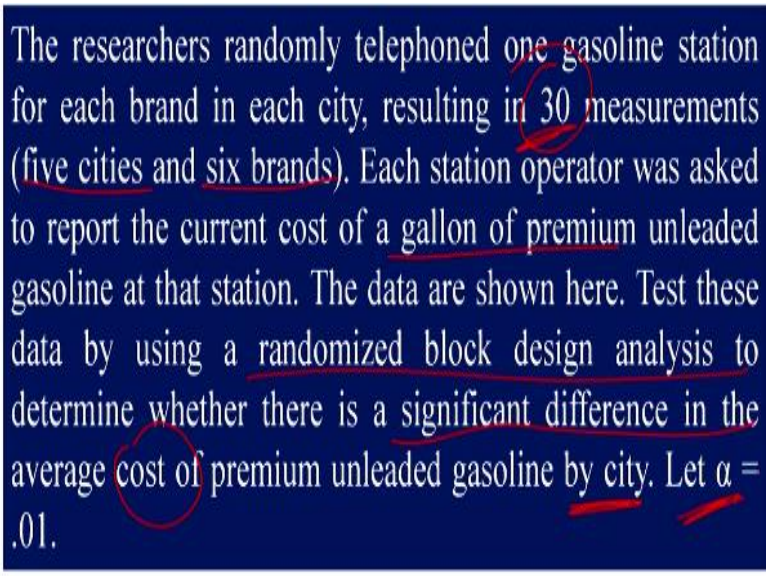
(Refer Slide Time: 25:07)



Let us look at one more example suppose, a national travel association studied the cost of premium unleaded gasoline in United States during summers of 2010, so from the experience the director of the association believes that there was a significant difference in the average cost of the gasoline among urban areas in different parts of the country, so here things that the price of the gasoline is different in different parts of the country.

So to test this belief they placed a random call to gasoline stations in 5 different cities in the researchers are they realise that the brand of gasoline might make a difference in price, so there independent variable is weather there is difference in cost of gasoline in different parts of the country or these 5 cities of the country and they were; they knew that the brand also affected price.

(Refer Slide Time: 26:39)



So, they set up an experiment and they got this particular data, so this how they collected data, they randomly you know called the gasoline stations and there were 5 different brands of gasoline across 6 different; there were 6 brands of gasoline's across 5 cities, right, so in all they collected data from 30 gasoline stations where in they chose one brand and one city, right so that is how they chose total; they collected data from 30 different gasoline stations.

So, each station operator was asked to report the current cost of the gasoline right, so the data are shown in the next slide, so this the test these data by using a randomised block design analysis to determine whether there is a significant difference in average cost of the gasoline by city, so there are 5 cities and gasoline prices were noted in all those cities for 6 brands, right so we need to test this hypothesis at 99% significance level, right.

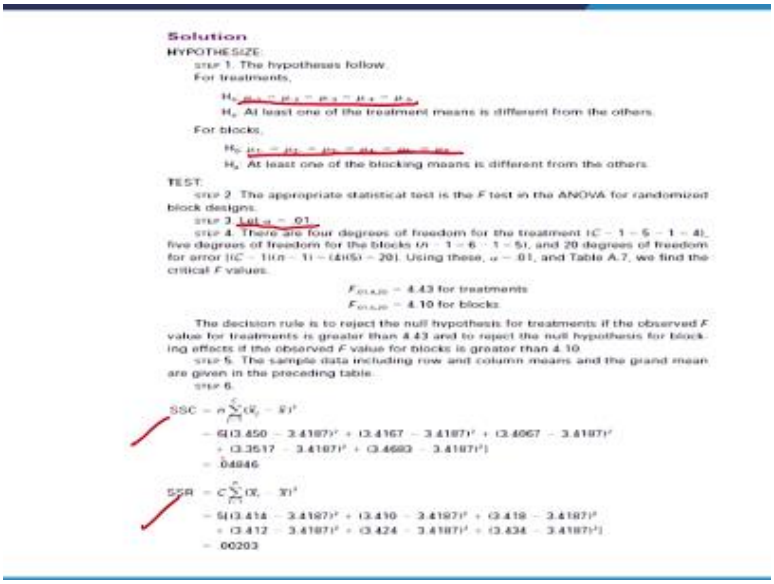
(Refer Slide Time: 28:03)

	Geographic Region				
Brand	Miami	Philadelphia	Minneapolis	San Antonio	Oakland
A	3.47	3.40	3.38	3.32	3.50
B	3.43	3.41	3.42	3.35	3.44
C	3.44	3.41	3.43	3.36	3.45
D	3.46	3.45	3.40	3.30	3.45
E	3.46	3.40	3.39	3.39	3.48
F	3.44	3.43	3.42	3.39	3.49

So, these the data, right, so these are different regions; Miami, Philadelphia and so on, right, Oakland, these are 6 different brands of gasoline, right so this is the price of brand A in Miami city, this is the price of brand F in Oakland right, so we need to test whether these means are same or not, right, so let us say Xm, right x bar m, x bar p for Philadelphia, right x bar o, right, so all these means are same or not, so null hypothesis is they are same, right.

An alternative hypothesis is not same, right, similarly the second null hypothesis is whether these means are same or not, the row means are same or not, there are six row means, right.

(Refer Slide Time: 29:04)



Solution

HYPOTHESIZE:

STEP 1. The hypotheses follow.

For treatments,

H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 = \mu_6$

H_a: At least one of the treatment means is different from the others.

For blocks,

H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 = \mu_6$

H_a: At least one of the blocking means is different from the others.

TEST:

STEP 2. The appropriate statistical test is the F test in the ANOVA for randomized block designs.

STEP 3. Let $\alpha = .01$.

STEP 4. There are four degrees of freedom for the treatment ($C - 1 = 5 - 1 = 4$), five degrees of freedom for the blocks ($n - 1 = 6 - 1 = 5$), and 20 degrees of freedom for error [$(C - 1)(n - 1) = (4)(5) = 20$]. Using these, $\alpha = .01$, and Table A.7, we find the critical F values:

$F_{.01,4,20} = 4.43$ for treatments

$F_{.01,5,20} = 4.10$ for blocks

The decision rule is to reject the null hypothesis for treatments if the observed F value for treatments is greater than 4.43 and to reject the null hypothesis for blocking effects if the observed F value for blocks is greater than 4.10.

STEP 5. The sample data including row and column means and the grand mean are given in the preceding table.

STEP 6.

$$SSC = n\sum_{j=1}^{c}(\bar{x}_j - \bar{\bar{x}})^2$$

$$= 6[(3.450 - 3.41167)^2 + (3.4167 - 3.41167)^2 + (3.4067 - 3.41167)^2 + (3.3517 - 3.41167)^2 + (3.4683 - 3.41167)^2]$$

$$= .04846$$

$$SSR = c\sum_{i=1}^{r}(\bar{x}_i - \bar{\bar{x}})^2$$

$$= 5[(3.414 - 3.41167)^2 + (3.410 - 3.41167)^2 + (3.419 - 3.41167)^2 + (3.412 - 3.41167)^2 + (3.424 - 3.41167)^2 + (3.434 - 3.41167)^2]$$

$$= .00203$$

So, you can solve this example similar to what we solved the earlier example, so you have got null hypothesis is that all these five column means are same, then for blocks or for blocking variable, all the six row means are same, right, so we need to test this hypothesis at .01 significance level, right, so after calculating the column sum of square, row sum of square.

(Refer Slide Time: 29:39)

$$SSE = \sum_{i=1}^{n}\sum_{j=1}^{c}(x_{ij} - \bar{x}_j - \bar{x}_i + \bar{x})^2$$

$$= (3.47 - 3.450 - 3.414 + 3.4187)^2 + (3.43 - 3.450 - 3.410 + 3.4187)^2 + \ldots + (3.48 - 3.4683 - 3.424 + 3.4187)^2 + (3.49 - 3.4683 - 3.434 + 3.4187)^2 = .01281$$

$$SST = \sum_{i=1}^{n}\sum_{j=1}^{c}(x_{ij} - \bar{x})^2$$

$$= (3.47 - 3.4187)^2 + (3.43 - 3.4187)^2 + \ldots + (3.48 - 3.4187)^2 + (3.49 - 3.4187)^2 = .06330$$

$$MSC = \frac{SSC}{C-1} = \frac{.04846}{4} = .01213$$

$$MSR = \frac{SSR}{n-1} = \frac{.00203}{5} = .00041$$

$$MSE = \frac{SSE}{(C-1)(n-1)} = \frac{.01281}{20} = .00064$$

$$F = \frac{MSC}{MSE} = \frac{.01213}{.00064} = 18.95$$

Source of Variance	SS	df	MS	F
Treatment	.04846	4	.01213	18.95
Block	.00203	5	.00041	
Error	.01281	20	.00064	
Total	.06330	29		

And let us say error sum of a square, you can find out total sum of a square and this is your F value, so F is you have to compare this F value with the critical value, right, okay.

(Refer Slide Time: 29:58)

Minitab Output

Two-way ANOVA: Gas Prices versus Brand, City

Source	DF	SS	MS	F	P
Brand	5	0.0020267	0.0004053	0.63	0.677
City	4	0.0485133	0.0121283	18.94	0.000
Error	20	0.0128067	0.0006403		
Total	29	0.0633467			

Excel Output

Anova: Two-Factor Without Replication

ANOVA

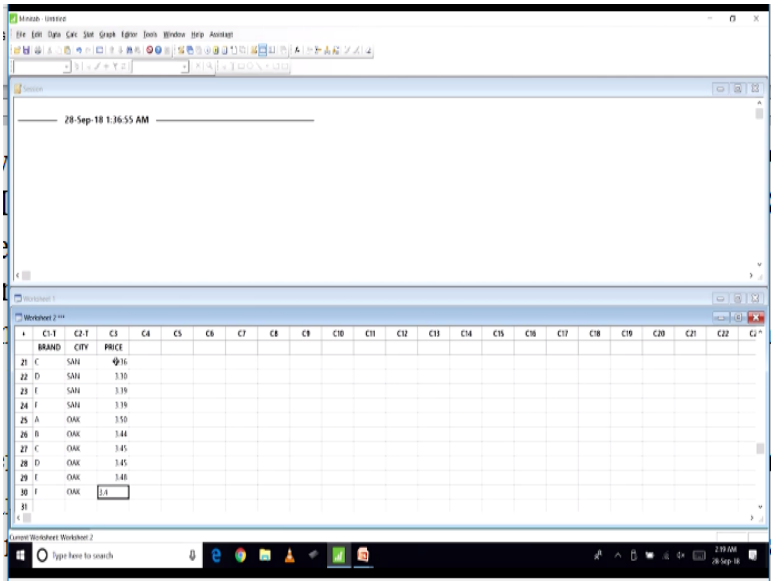
Source of Variation	SS	df	MS	F	P-value	F crit
Rows	0.002027	5	0.000405	0.63	0.6768877	4.10
Columns	0.048513	4	0.012128	18.94	0.0000014	4.43
Error	0.012807	20	0.000640			
Total	0.063347	29				

So, let us find out answer to this question and this is our Minitab output, so let us look at what was the F; what was the significance level, it was .01, right, so .01 alpha is .01 right, 0.01, so p

value is this, so we will reject null hypothesis this, it means city wise there is a significant difference in price but brand wise, there is no significant difference because you will not reject null hypothesis here and here you will reject null hypothesis.

So, we will find out solution to this particular question, in fact this is an output from Excel, this ANOVA table, so let us solve this question using Minitab.

(Refer Slide Time: 30:59)



So, we can go for data entry, first of all, so you can have a new file, so we will enter data for all these 30 observations, so you have got brand city and price; price is nothing but your response variable or dependent variable, right, so brand A, B, C, D, E and then F, right, so you can copy all these, so this is you will have 30 observations, right and one more time, right so, this is you have got observation for the entry for brands and you have got cities, right.

So, if you want let it be C1 C2 C3 C4 C5 right rather than writing, okay, Philadelphia, okay let it be PHILA, right, so up to 12 right, the next is the next city, so up to 18, this will go up to 18, then next city up to 24 and finally, Oakland, right and then you can write prices, so you can just copy and paste it up to point number 30, right or serial number 30, let us write prices over here, so brand A Miami 3.47, then brand B first city, brand C, brand D, brand E and last brand is F brand, right.

So, there are total six brands, so these are the prices of six brands in Miami city, then price of the six brands in Philadelphia, so this is brand A, the third city 3.38, then 3.42, brand C then D, E and F, right, now we are left with 2 more cities, right for all the six brands, so once you are done with these data entry, you are supposed to look at the ANOVA table and you need to look at p value specifically.

Because there you are deciding whether you are rejecting or not rejecting null hypothesis, so this is how you should go for data entry and then go to stat, go to ANOVA, it is general linear model, fit general linear model, so your price is now response variable and the factors are brand and city, right, select, you need to look at the significance level it is correct, it is .99 or 99%, so we will click it okay.

So, we will go to ANOVA table and see what are the values of the p, right, so look at the same answer, right, what I had shown in slides just I will show you once again, yeah just see, it is .667 and 0, right, so you will reject null hypothesis, and you will say that there is a significant difference in price so for as cities are concern but there is no affect or there is no difference in price as far as brands are concern, so with this let me complete today's session, in next session we will have some more examples, thank you.