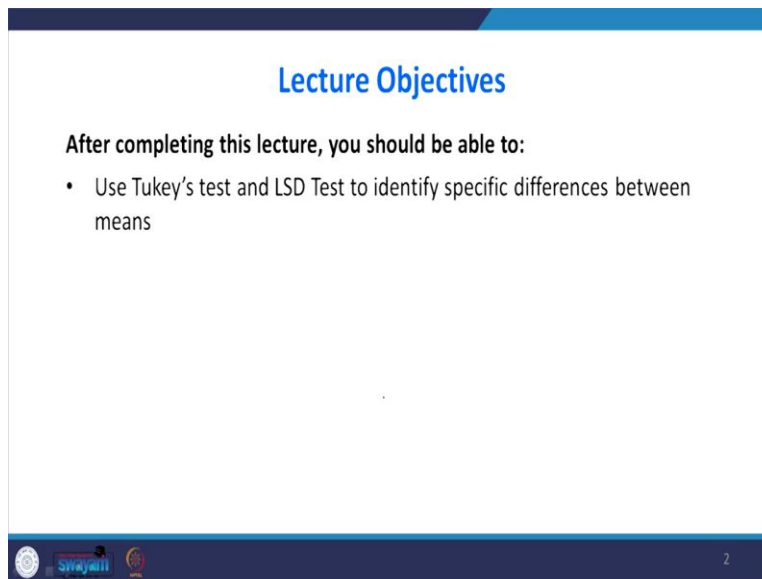


Data Analytics with Python
Prof. Ramesh Anbanandam
Department of Management Studies
Indian Institute of Technology – Roorkee

Lecture – 25
Post Hoc Analysis(Tukey's test)

Dear students in the previous class we have seen the theoretical background behind analysis of variance and also we have solved a problem. So, what will happen once you reject a null hypothesis of an ANOVA problem that means you are accepting alternative hypothesis it is need not be that all means not equal some time any pair may be equal some other pair may be not equal. So, whenever you reject a null hypothesis then we have to say which two means are equal for that purpose there is one more statistical analysis that is a Post Hoc analysis useful one test is Tukey Kramer test another test is HSD test. We will see with that how to use this post hoc analysis in ANOVA in this lecture.

(Refer Slide Time: 01:22)



Lecture Objectives

After completing this lecture, you should be able to:

- Use Tukey's test and LSD Test to identify specific differences between means

swayamii

So, the lecture objective is after completing the lecture you should be able to use Tukey test and least that is LST test to identify specific differences between means.

(Refer Slide Time: 01:34)

Designing engineering experiments

- Experimental design methods are also useful in engineering design activities, where new products are developed and existing ones are improved
- By using designed experiments, engineers can determine which subset of the process variables has the greatest influence on process performance



3

We will take in this problem and engineering perspective experimental design methods also useful in engineering design activities where new products are developed and existing ones are improved. By you see design of experiments engineers can determine which subset of the process variables has the greatest influence on the process performance that is the main objective of the design of experiment is. What kind of variables that has the greatest influence on the performance of the product.

(Refer Slide Time: 02:09)

Designing engineering experiments

- The results of an experiment can lead to
 1. Improved process yield
 2. Reduced variability in the process and closer conformance to nominal or target requirements
 3. Reduced design and development time
 4. Reduced cost of operation

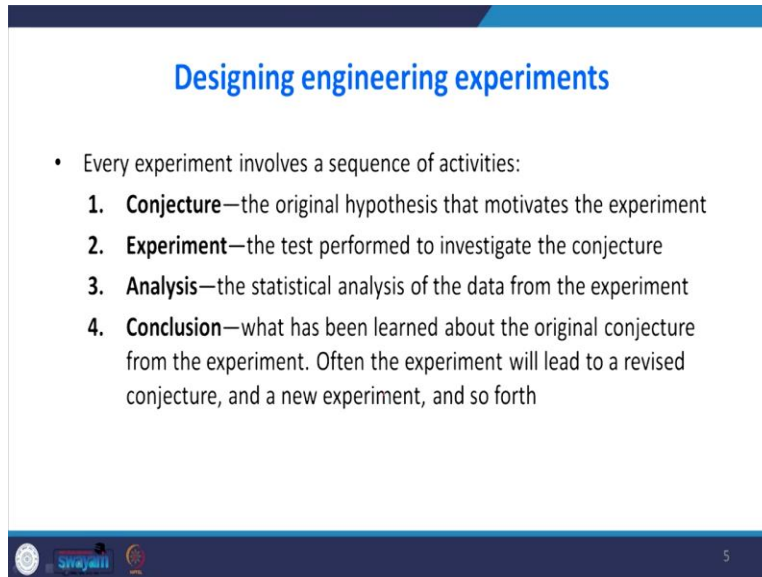


4

When you do the design of experiments what are the advantages benefits one is it improves the process yield it reduces the variability in the process that leads to less rejection and closer confirmation to the nominal or target requirements, so quality of the product is improved and

reduced design and development time because before making the product since we are doing the design experiments so the time spent on redesign is reduced at the same time reduced the cost of operations because the waste is minimized.

(Refer Slide Time: 02:48)



Designing engineering experiments

- Every experiment involves a sequence of activities:
 1. **Conjecture**—the original hypothesis that motivates the experiment
 2. **Experiment**—the test performed to investigate the conjecture
 3. **Analysis**—the statistical analysis of the data from the experiment
 4. **Conclusion**—what has been learned about the original conjecture from the experiment. Often the experiment will lead to a revised conjecture, and a new experiment, and so forth

5

Some of the term that we have to remember while going for this design of experiment is one does a conjecture, Conjecture is the original hypothesis that motivates the experiment, experiment the tests performed to investigate the conjecture. Analysis the statistical analysis of the data from the experiment, so, conclusion is what has been learned about the original conjecture from the experiment often the experiment will lead to a revised conjecture and new experiment and so forth.

(Refer Slide Time: 03:23)

The completely randomized single-factor experiment example

- A manufacturer of paper that is used for making grocery bags is interested in improving the tensile strength of the product
- Product engineer thinks that tensile strength is a function of the hardwood concentration in the pulp and that the range of hardwood concentrations of practical interest is between 5 and 20%.



Reference: Applied statistics and probability for engineers, Douglas C. Montgomery, George C. Runger, John Wiley & Sons, 2007



6

We will solve one, one way problem in this class then I will explain how to use post hoc analysis. The problem is like this a manufacturer of paper industry he is using the paper for making grocery bags and he want to improve the tensile strength of the product. The product engineer thinks that the tensile strength is a function of hardwood concentration in the pulp and that the range of hardwood concentration is concentration of practical interest is between 5 to 20% so what is the meaning is that when you increase the hardwood concentration so the tensile strength will increase.

A team of engineers responsible for the study decides to investigate for level of hardwood concentrations the concentration level which they considers our 5% age 10% 15% and 20% they decide to make up 6 test specimen at each concentration level using a pilot plant all 24 specimens are tested on a laboratory tensile tester in a random order the data from this experiment is shown in the table.

(Refer Slide Time: 04:40)

The completely randomized single-factor experiment example

- Tensile Strength of Paper (psi)

Hardwood Concentration (%)	Observations						Total	Avg
	1	2	3	4	5	6		
5	7	8	15	11	9	10	60	10.00
10	12	17	13	18	19	15	94	15.67
15	14	18	19	17	16	18	102	17.00
20	19	25	22	23	18	20	127	21.17
							383	15.96

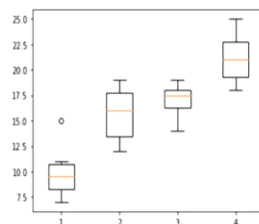
Now the row 5 represents hardwood concentration 5% 10% 15% and 20% 5% 10% these are the observations there are 6 times the experiment is repeated the average value is given so here the treatment is the percentage of hardwood concentration.

(Refer Slide Time: 05:02)

The completely randomized single-factor experiment example

```
In [3]: fivepercent = [7,8,15,11,9,10]
tenpercent = [12,17,13,18,19,15]
fifteenpercent = [14,18,19,17,16,18]
twentypercent = [19,25,22,23,18,20]

box_plot_data = [fivepercent, tenpercent, fifteenpercent, twentypercent]
plt.boxplot(box_plot_data)
plt.show()
```



First we will plot it with help of box and whisker plot so I am going to take the first data set as a 5% in array 5 8 15 11 9 10, next variable name is a 10 % I am ever taken the next array 12 7 13 18 19 15 then 15% and I have taken all the 6 variables in 15% in 20% and so on. So, I go to draw the box plot, so box underscore plot underscore data so just to call that arrays for 5% 10% 15% and 20% then you write plt dot box plot box underscore plot underscore data plot dot show, so we are getting the box and whisker plot.

So what is happening here you see the means are not equal there is a lot of differences they there appears that whenever the percentage of hardwood is increasing so the tensile strength is increasing because there seems to be there is an increasing strength.

(Refer Slide Time: 06:13)

Typical Data for Single Factor Experiment

Treatment	Observations				Totals	Averages
1	y_{11}	y_{12}	...	y_{1n}	$y_{1.}$	\bar{y}_1
2	y_{21}	y_{23}	...	y_{2n}	$y_{2.}$	\bar{y}_2
.
.
.
a	y_{a1}	y_{a2}	...	y_{an}	$y_{a.}$	\bar{y}_a
					$y_{..}$	$\bar{y}_{..}$

This is a typical data for single factor experiment generally see the treatment is taken as 1 2 3 4 observations are taken in row wise y_{11} this is the response variable of first a treatment and first a sample first treatment second sample first treatments and $y_{1.}$ $y_{n.}$ so if I write $y_{1.}$ that is a total the first row total $y_{2.}$ dot second row total. If I write a is there is a a treatment a levels is it so if I write $y_{a.}$ the dot represents the sum so $y_{a.}$ dot means the totals.

If I write \bar{y}_1 dot bar so that is the row 1 mean \bar{y}_2 dot bar that is the second row mean by a dot ath level averages. If you write $y_{..}$ dot, dot does the sum of all total if I write $\bar{y}_{..}$ dot dot bar that is the average.

(Refer Slide Time: 07:17)

Sum of Squares

$$\text{Total sum of squares} = SS_1 = \sum_{i=1}^a \sum_{j=1}^n (y_{ij} - \bar{y}_{..})^2$$

$$\text{Treatment sum of squares} = SS_{\text{Treatments}} = n \sum_{i=1}^a (\bar{y}_{i.} - \bar{y}_{..})^2$$

$$\text{Error sum of Squares} = SSE = \sum_{i=1}^a \sum_{j=1}^n (y_{ij} - \bar{y}_{j.})^2$$

So, SST treatment sum of square is the formula which you have seen previously because if you write in this way for especially for as a student this will be an easy way to solve the problem in the examination. So, this will save a lot of time. $\sum_{i=1}^a$ equal to 1 to a up to level j equal to 1 to n up to number of sample size y_{ij} is the individual response - $\bar{y}_{..}$ but that is overall mean so that will give the total sum of square.

If you if you want to move the treatment sum of square $SS_{\text{Treatment}}$ equal to n into $\bar{y}_{i.} - \bar{y}_{..}$ that is over all this is the row 1 mean this is overall mean whole square multiplied by n number of sample in the row 1 y . SSE is $y_{ij} - \bar{y}_{j.}$ that corresponding mean whole square okay this is your error sum of square.

(Refer Slide Time: 08:25)

ANOVA with Equal Sample Sizes

$$SST = \sum_{i=1}^a \sum_{j=1}^n y_{ij}^2 - \frac{y^2_{..}}{N}$$

$$SSTreatments = \frac{1}{n} \sum_{i=1}^a y_{i.}^2 - \frac{y^2_{..}}{N}$$

$N = an =$ No. of Treatments \times no. of sample size = Total no. of Sample Size

There is a shortcut formula for this the shortcut formula because this formula is very useful if you are using calculator. So, when you simplify the previous slide with what you are done so SST is nothing but Sigma of I equal to 1 to a sigma j equal to 1 to n y ij square - y dot dot squared order by n, n is the nothing but a into n is the number of level n is small n is number of observation at level a. So, trick this is total sum of square the treatment sum of square is 1 by n Sigma i equal to 1 to a y i dot m square - y dot dot whole square.

When you see that here also y dot dot square by n here also y dot square n is same so if you calculate for 1 it can be used for both calculation. So, we know that SST equal to SSTreatment plus SSE so when you subtract it you can get SSE, so this will save a lot of time in the examination. The previously there is a equal sample size if there is unequal sample size yes as T is same y IJ square - y dot square by n but the SS Treatment is yi dot squared order by n i because this this new term will be introduced there - y double dot whole square divided by n.

(Refer Slide Time: 09:51)

Problem: Analysis of variance

- Consider the paper tensile strength experiment described.
- We can use the analysis of variance to test the hypothesis that different hardwood concentrations do not affect the mean tensile strength of the paper.
- The hypotheses are
- $H_0: \tau_1 = \tau_2 = \tau_3 = \tau_4 = 0$
- $H_1: \tau_i \neq 0$ for at least one i



14

Consider the paper tensile strength experiment which is described previously. We can use the analysis of variance to test hypothesis that the different hardwood concentration do not affect the mean tensile strength of the paper. So, what is the null hypothesis is that that different hardwood concentration does not affect the mean tensile strength that means the hardwood concentration tensile since are independent nothing to do with that one.

So, they are here the hypothesis is different treatment effective 0 the alternative hypothesis of this is there is a effect of treatment.

(Refer Slide Time: 10:30)

Problem: Analysis of variance

- We will use $\alpha = 0.01$.
- The sums of squares for the analysis of variance are computed are as follows:

$$\begin{aligned}SS_T &= \sum_{i=1}^4 \sum_{j=1}^6 y_{ij}^2 - \frac{y_{..}^2}{N} \\ &= (7)^2 + (8)^2 + \dots + (20)^2 - \frac{(383)^2}{24} = 512.96 \\ SS_{\text{Treatments}} &= \sum_{i=1}^4 \frac{y_i^2}{n} - \frac{y_{..}^2}{N} \\ &= \frac{(60)^2 + (94)^2 + (102)^2 + (127)^2}{6} - \frac{(383)^2}{24} = 382.79 \\ SS_E &= SS_T - SS_{\text{Treatments}} \\ &= 512.96 - 382.79 = 130.17\end{aligned}$$



15

When we take alpha equal to 1% the sum of square of analysis of variance are computed as follows we can say $\sum y_{ij}^2 - \frac{(\sum y_i)^2}{n}$ we can simplify we just you can substitute this formula so SS treatment is 512.96 sorry total sum of square SS treatment is 382.79 when you subtract it will get SSE so SSE is $512.96 - 382.79$ it is 130.17.

(Refer Slide Time: 11:01)

Problem: Analysis of variance

- The ANOVA is summarized as follow

Source of Variation	Sum of Squares	Degrees of freedom	Mean Square	F ₀	P-value
Hardwood concentration	382.79	3	127.6	19.6	3.59 E-6
Error	130.17	20	6.51		
Total	512.96	23			

```

In [9]: from scipy import stats
1-scipy.stats.f.cdf(19.6, 3, 20)
Out[9]: 3.599599239012541e-06

```

So, this is an ANOVA setup when you supply this value here so you can get what you are done we got SST we got SS treatment when you subtract it will get SSE error so degrees of freedom is there is a totally 24 element so $24 - 1$, 23. So, there was a 4 rows, $4 - 1$, 3 so $23 - 3$ is 20 then when you divide this 382.9 degrees of freedom we will get 127.6 when you divide 172 divided by 20 will get 6.5. so, 127.62 by 6.5 you will be 19.6 so it look like 19.6 big, so what we can do so it said it will this is a calculated value.

So what is a table value so when F is 19.6 numerator degrees of freedom is 3 denominator degrees of freedom is 20, so when you subtract it so 1 minus so that will give you the p value p value is 3.59 into 10 to the power - 6 you see this values is very, very, very low. So, we are two because this p-value is less than alpha if you say alpha equal to 5% we have to reject null hypothesis when you reject a null hypothesis that there is the influence of this hardwood on the tensile strength.

(Refer Slide Time: 12:37)

Problem: Analysis of variance

- Since $f_{0.01,3,20} = 4.94$, we reject H_0 and conclude that hardwood concentration in the pulp significantly affects the mean strength of the paper

```
In [32]: scipy.stats.f.ppf(1-0.01, dfn=3, dfd=20)
```

```
Out[32]: 4.938193382310539
```



18

Since $F_{1\% 3, 20}$ is 4.94 so when you 3, 20 you see that when you suppose if you are comparing the critical value so this value is 4.9 which you got from this one you can see that `scipy dot stats dot f dot ppf`, if it is a 1% so we want to know probability of when the right side area is 0.01 so for that 1 `scipy dot statistical F dot 1 - 0.01` will give you the because we want to know the right side area but the Python gives only the left side area so `one - 0.01` that probability when degrees of freedom is 3 when the denominator is no 20 the corresponding F value is 4.93. Our calculated F value is 19.6, so 19.6 far away from gates so we got to reject a null hypothesis.

(Refer Slide Time: 13:43)

Problem: Analysis of variance

```
In [23]: scipy.stats.f_oneway(fivepercent,tenpercent,fifteenpercent,twentypercent)
```

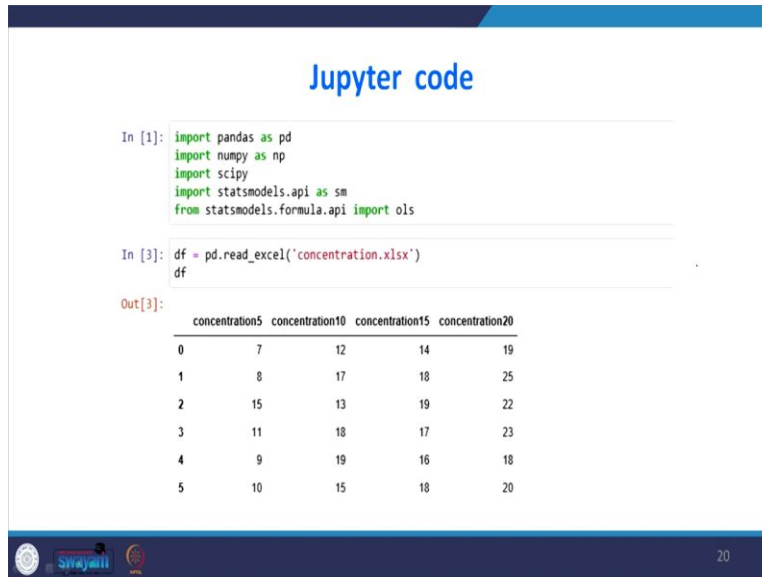
```
Out[23]: F_onewayResult(statistic=19.605206999573184, pvalue=3.5925782584743027e-06)
```



19

So, what we can do we can directly we can run ANOVA so cipher dot starts dot F underscore one way so call 5% 10% 15% 20% so you are getting F statistics 19.68 CP value is 3.59 into 10 to power – 6, so both the way you are getting the same result.

(Refer Slide Time: 14:06)



The screenshot shows a Jupyter Notebook interface with the title "Jupyter code". It contains two input cells and one output cell. The first input cell shows the following code: `import pandas as pd`, `import numpy as np`, `import scipy`, `import statsmodels.api as sm`, and `from statsmodels.formula.api import ols`. The second input cell shows `df = pd.read_excel('concentration.xlsx')` and `df`. The output cell shows a DataFrame with 6 rows and 4 columns: concentration5, concentration10, concentration15, and concentration20. The data values are as follows:

	concentration5	concentration10	concentration15	concentration20
0	7	12	14	19
1	8	17	18	25
2	15	13	19	22
3	11	18	17	23
4	9	19	16	18
5	10	15	18	20

Now we will solve this problem suppose assume that this dataset which I have already entered into the excel file. So, import pandas as pd import numpy as np import scipy import stats model dot a p dot sm from stats model dot Formula d api import ols so d of equal to p d dot read underscore excel I saved that file name is concentration constantly.xlsx. So, when you write this df you are getting so concentration 1 concentration 2 concentration 3 consultation for that is a different % level.

Now we have to convert this one into only in two column in one column I need to have concentration level in our another column I am going to have only the values response variables.

(Refer Slide Time: 14:56)

Jupyter code

```

In [5]: data_r1 = pd.melt(df.reset_index(), id_vars='index', value_vars=['concentration5', 'concentration10', 'concentration15', 'concentration20'])
data_r1.columns = ['index', 'treatments', 'value']

In [6]: model = ols('value ~ C(treatments)', data=data_r1).fit()

In [7]: model.summary()

```

For that I have to use melt function so I am going to save that file in the object called data underscore r 1 pd dot melt df dot reset underscore index, id underscore vars equal to index, value underscore vars equal to concentrations file that is the value of the variables concentration of 5 concentration 10 concentration 15 concentration 20 so data underscore r 1 columns going to be index treatment and values. So, model ols values tilde c in bracket treatments data equal to data underscore r 1 dot fit. So when I write model dot summary I will be getting this result right. **(Refer Slide Time: 15:44)**

Jupyter code

Out[7]: OLS Regression Results

Dep. Variable:	value	R-squared:	0.746
Model:	OLS	Adj. R-squared:	0.700
Method:	Least Squares	F-statistic:	19.61
Date:	Tue, 27 Aug 2019	Prob (F-statistic):	3.59e-06
Time:	15:03:38	Log Likelihood:	-54.344
No. Observations:	24	AIC:	116.7
Df Residuals:	20	BIC:	121.4
Df Model:	3		
Covariance Type:	nonrobust		

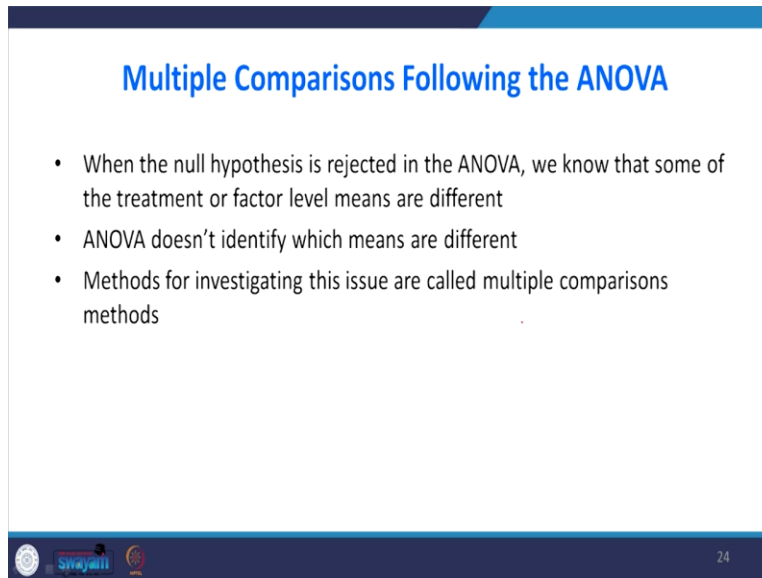
	coef	std err	t	P> t	[0.025	0.975]
Intercept	15.6667	1.041	15.042	0.000	13.494	17.839
C(treatments)[T.concentration15]	1.3333	1.473	0.905	0.376	-1.739	4.406
C(treatments)[T.concentration20]	5.5000	1.473	3.734	0.001	2.428	8.572
C(treatments)[T.concentration5]	-5.6667	1.473	-3.847	0.001	-8.739	-2.594

Omnibus: 0.929 Durbin-Watson: 2.181
Prob(Omnibus): 0.628 Jarque-Bera (JB): 0.861
Skew: 0.248 Prob(JB): 0.650
Kurtosis: 2.215 Cond. No. 4.79

So what we are getting here this was the regression model now we are getting the ANOVA table so aov underscore table equal to some dot stat dot anova underscore alum when you call that model type equal to one you type this table you will get treatment there are 4 row was there so 4

- 13 degrees of freedom for the error degrees of freedom is 20 so this is treatment sum of square error sum of square when you divide by 3 you will get 127.59 when you divide 130 by 20, 6.59 so when you divide 127.59 into 509 this is your calculated values the p-value is very, very low. So, we can reject the null hypothesis.

(Refer Slide Time: 16:37)



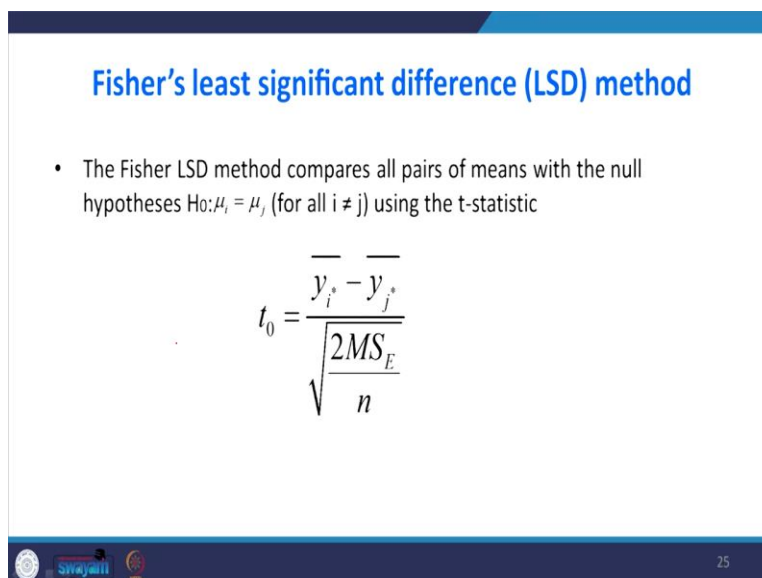
Multiple Comparisons Following the ANOVA

- When the null hypothesis is rejected in the ANOVA, we know that some of the treatment or factor level means are different
- ANOVA doesn't identify which means are different
- Methods for investigating this issue are called multiple comparisons methods

24

When null hypothesis rejected in the ANOVA we know that some of the treatment or factor level means are different. Anova does not identify which means are different methods for investigating this issue is called multiple comparison method are post hoc analysis that we will do here.

(Refer Slide Time: 16:57)



Fisher's least significant difference (LSD) method

- The Fisher LSD method compares all pairs of means with the null hypotheses $H_0: \mu_i = \mu_j$ (for all $i \neq j$) using the t-statistic

$$t_0 = \frac{\bar{y}_i - \bar{y}_j}{\sqrt{\frac{2MS_E}{n}}}$$

25

One technique for doing post hoc analysis Fisher's least significant difference method, the Fisher's least significant difference method compares all pairs of mean with the null hypothesis $H_0: \mu_i = \mu_j$ for all $i \neq j$ using the t statistics this is nothing but your two sample t-test. So, $y_i - y_j$ divided by root of $\frac{2MSE}{n}$ we got this one generally we will get this one $\frac{2MSE}{n}$ greater by n both are same we wrote 2 into $\frac{2MSE}{n}$ divided by n but here is the sample size is same.

(Refer Slide Time: 17:56)

Fisher's least significant difference (LSD) method

- Assuming a two-sided alternative hypothesis, the pair of means i and j would be declared significantly different if

$$|\bar{y}_i - \bar{y}_j| > LSD$$

where LSD, the least significant difference, is

$$LSD = t_{\alpha/2, a(n-1)} \sqrt{\frac{2MSE}{n}}$$

26

Assuming a two-sided alternative hypothesis the pair of means i and j would be declared significantly different if the absolute value of the different of their mean if it is greater than LSD then we will say that there is a significant difference between that 2 pair, so LSD if you bring the left hand side the previous formula will be $t_{\alpha/2, a(n-1)} \sqrt{\frac{2MSE}{n}}$ a is the number of levels in this number of observations each treatment minus MSE divided by n just I have readjusted that form.

(Refer Slide Time: 18:34)

Fisher's least significant difference (LSD) method

- If the sample sizes are different in each treatment, the LSD is defined as

$$LSD = t_{\alpha/2, N-a} \sqrt{MS_E \left(\frac{1}{n_i} + \frac{1}{n_j} \right)}$$

So if the sample sizes are different in each treatment the LSD is defined as LSD equal to t alpha by 2 n - a root of MSE by 1 divided by n i + 1 divided by n j that means for each pair there will be a different LSD because the sample size are different that should be very careful on that one.

(Refer Slide Time: 19:00)

Problem : LSD method

- We will apply the Fisher LSD method to the hardwood concentration experiment. There are a = 4 means, n = 6, MSE = 6.51, and $t_{0.025, 20} = 2.086$.

The treatment means are

$$\bar{y}_{1.} = 10.00 \text{ psi}$$

$$\bar{y}_{2.} = 15.67 \text{ psi}$$

$$\bar{y}_{3.} = 17.00 \text{ psi}$$

$$\bar{y}_{4.} = 21.17 \text{ psi}$$

We will apply the Fisher LSD method to the hardwood concentration experiment there are 4 means n equal to 6 MS is 6.51 so in alpha equal to 5% for 0.025 and 20 degrees of freedom the t value is 2.086 this was the mean of different treatment.

(Refer Slide Time: 19:24)

Problem : LSD method

- The value of LSD is:

$$LSD = t_{0.025, 20} \sqrt{\frac{2MS_E}{n}} = 2.086 \sqrt{\frac{2(6.51)}{6}} = 3.07$$

- Therefore, any pair of treatment averages that differs by more than 3.07 implies that the corresponding pair of treatment means are different.

So, when you substitute here LSD value is 3.07 so we have to compare see that this one 1 and 2, 1 and 3, 1 and 4 - 1 3 and 2 and 4 therefore any pair of treatment averages that differs by more than 3.07 implies that that corresponding pair of treatment means are different. So, what you have to do next step or to take any two pair of the mean you have to find their absolute difference if the absolute difference is greater than 3.07 we can conclude that that two pairs means are different.

This was already we have got this ANOVA table now we will go for this LSD test import mat first we will find out the t value t values - 1 x scipy dot stats dot t dot ppf of 0.025, 20 because why I am taking - 1 because whether it is the right side value the t value should be positive. So, n equal to 6 MSE is already we know that it is this value 6.50 so LSD is I am writing this formula t multiplied by math dot square root of 2 MSE divided by n we are getting 3.07. So this value we got it already 3.07 this 3.07.

(Refer Slide Time: 20:49)

Problem : LSD method

- The comparisons among the observed treatment averages are as follows:

$$4 \text{ vs. } 1 = 21.17 - 10.00 = 11.17 > 3.07$$

$$4 \text{ vs. } 2 = 21.17 - 15.67 = 5.50 > 3.07$$

$$4 \text{ vs. } 3 = 21.17 - 17.00 = 4.17 > 3.07$$

$$3 \text{ vs. } 1 = 17.00 - 10.00 = 7.00 > 3.07$$

$$3 \text{ vs. } 2 = 17.00 - 15.67 = 1.33 < 3.07$$

$$2 \text{ vs. } 1 = 15.67 - 10.00 = 5.67 > 3.07$$

Now we are going to take all the pair's first you will take 4 versus 1, 4 versus 2, 4 versus 3 then 3 versus 1, 3 versus 2 and 2 versus 1, so the absolute difference is 11.17, 5.50 this is greater than so what we can conclude $\mu_4 \neq \mu_1$ but here you see that if this is less than 3.07 so what we have to conclude is this $\mu_3 = \mu_2$ all other pairs are different.

(Refer Slide Time: 21:29)

Problem : LSD method

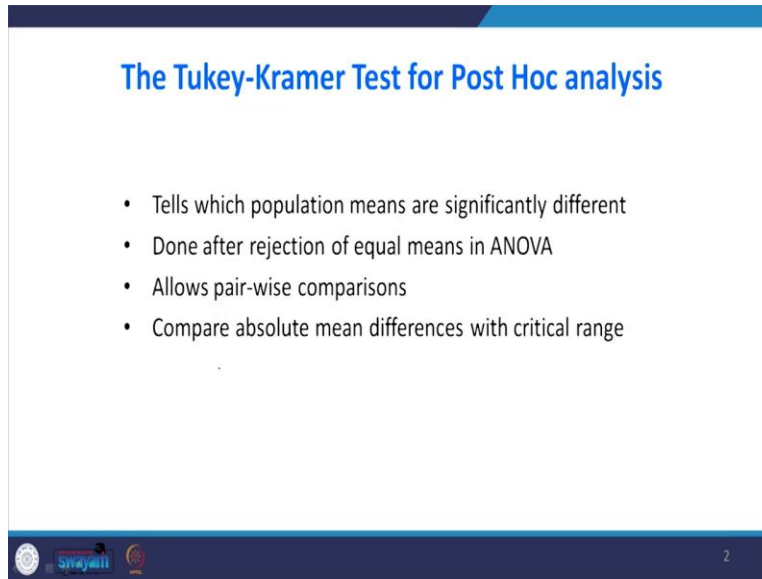
- In this problem we see that there are significant differences between all pairs of means except 2 and 3
- This implies that 10 and 15% hardwood concentration produce approximately the same tensile strength and that all other concentration levels tested produce different tensile strengths

In this problem we see that there are significant differences between all the pairs of mean except 2 and 3 this implies that 10 % and 15% hardwood concentration produce approximately the same tensile strength there is means are equal that two means are equal. So, what we are concluding that 10% and 15 % hardwood concentration produce approximately the same tensile strength and that all other concentration levels tested produce different tensile strength.

(Refer Slide Time: 11:01)

The Tukey-Kramer Test for Post Hoc analysis

- Tells which population means are significantly different
- Done after rejection of equal means in ANOVA
- Allows pair-wise comparisons
- Compare absolute mean differences with critical range

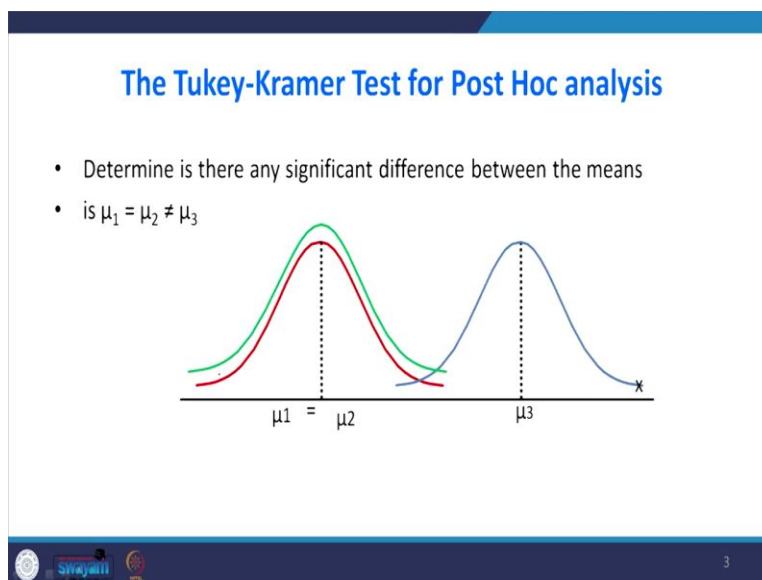


We will go for another post hoc analysis that that is called to Tukey Kramer test Tukey Kramer test tells which population means are significantly different it is then after rejection of equal means in ANOVA that means after rejecting our hypothesis it allows pairwise comparison all the means are compared as a pair. So, compare absolute mean differences with the critical range what will happen do in this taste test we will find out mean absolute difference so that difference is compared with the critical range that we got from the table called Tukey table.

(Refer Slide Time: 22:40)

The Tukey-Kramer Test for Post Hoc analysis

- Determine is there any significant difference between the means
- is $\mu_1 = \mu_2 \neq \mu_3$



So, Tukey Kramer test for post hoc analysis determine if there is any significant difference between the means so when we reject a null hypothesis this figure says that mu 1 equal to mu 2

but μ_3 is different. So, this which two pairs of means is equal that we can find with the help of this Tukey Kramer test.

(Refer Slide Time: 23:05)

Tukey-Kramer Critical Range

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

where:

- Q_U = Value from Studentized Range Distribution with c and $n - c$ degrees of freedom for the desired level of α
- MSW = Mean Square Within
- n_j and $n_{j'}$ = Sample sizes from groups j and j'

So, here we have to find out the critical range the critical range is Q_U root of MSW or we can say MS within the column otherwise we can say MSE divided by 2 $1/n_j + 1/n_{j'}$ divided by $n_j - n_{j'}$ and j dash is 2 pairs which are comparing and corresponding sample size. Here the Q_U the value from studentized range that I will show you I have the table with me studentized range distribution with the c and $n - c$ degrees of freedom.

Here c is the number of columns n is the total number of sample size degrees of freedom for the desired level of alpha MSW is mean square within nothing but every a MSE n_j and $n_{j'}$ says or sample sizes from groups j and j dash that means we are taking 2 pairs of population j and j dash they are comparing the we are finding the absolute difference. If that absolute difference is greater than critical range we will say that that two pairs are different. If it is within the critical range we say that that 2 pairs means is same.

(Refer Slide Time: 14:20)

Problem: Tukey- Kramer test

- Tensile Strength of Paper (psi)

Hardwood Concentration (%)	Observations						Total	Avg
	1	2	3	4	5	6		
5	7	8	15	11	9	10	60	10.00
10	12	17	13	18	19	15	94	15.67
15	14	18	19	17	16	18	102	17.00
20	19	25	22	23	18	20	127	21.17
							383	15.96

This was the problem which have solved previously so for this problem we have remember we have rejected null hypothesis.

(Refer Slide Time: 24:30)

The Tukey-Kramer Procedure

1. Compute absolute mean differences:

$$|\bar{x}_1 - \bar{x}_2| = |10.00 - 15.67| = 5.67$$

$$|\bar{x}_1 - \bar{x}_3| = |10.00 - 17.00| = 7$$

$$|\bar{x}_2 - \bar{x}_3| = |15.67 - 17.00| = 1.33$$

$$|\bar{x}_1 - \bar{x}_4| = |10.00 - 21.17| = 11.17$$

$$|\bar{x}_2 - \bar{x}_4| = |15.67 - 21.17| = 5.5$$

$$|\bar{x}_3 - \bar{x}_4| = |17.00 - 21.17| = 4.17$$

So first we are finding out reject a null hypothesis then we are going to do Tukey Kramer. So, we are going to compare mean of X 1 bar and X 2 bar, X 1 bar and X 3 bar X 2 bar and X 3 bar X 1 X 4 X 2 X 4 X 3 and X 4. So, the absolute mean 4 X 1 bar - X 2 bar is 5.67 for second one is 7 for third one is 1.33 X 1 and X 4 is 11.14 and so on.

(Refer Slide Time: 15:04)

The Tukey-Kramer Procedure

2. Find the Q_U value from the table with $c = 4$ and $(n - c) = (24 - 4) = 20$ degrees of freedom for the desired level of α ($\alpha = .05$ used here):

$$Q_U = 3.96$$

Next we will find out the critical range find QU value from the table with see here number of treatment is 4 n - c is 20 degrees of freedom for the desired allele of alpha equal to 5% this 3.9 c is got from this table.

(Refer Slide Time: 251:22)

Error Term	2	3	4	5	6	7	8	9	10
5	3.64 5.70	4.60 6.90	5.22 7.60	5.67 8.42	6.03 8.91	6.33 9.32	6.58 9.67	6.80 9.97	6.99 10.24
6	3.46 5.24	4.34 6.33	4.90 7.03	5.30 7.56	5.63 7.97	5.90 8.32	6.12 8.61	6.32 8.87	6.49 9.10
7	3.34 4.95	4.16 5.92	4.68 6.54	5.06 7.01	5.36 7.37	5.61 7.60	5.82 7.94	6.00 8.17	6.16 8.37
8	3.26 4.75	4.04 5.64	4.53 6.20	4.89 6.62	5.17 6.96	5.40 7.24	5.60 7.47	5.77 7.68	5.92 7.86
9	3.20 4.60	3.95 5.43	4.41 5.96	4.76 6.35	5.02 6.66	5.24 6.91	5.43 7.13	5.59 7.33	5.74 7.49
10	3.15 4.48	3.88 5.27	4.33 5.77	4.65 6.14	4.91 6.43	5.12 6.67	5.30 6.87	5.46 7.05	5.60 7.21
11	3.11 4.39	3.82 5.15	4.26 5.62	4.57 5.97	4.82 6.25	5.03 6.48	5.20 6.67	5.35 6.84	5.49 6.99
12	3.08 4.32	3.77 5.05	4.20 5.50	4.51 5.84	4.75 6.10	4.95 6.32	5.12 6.51	5.27 6.67	5.39 6.81
13	3.06 4.26	3.73 4.96	4.15 5.40	4.45 5.73	4.69 5.98	4.88 6.19	5.05 6.37	5.19 6.53	5.32 6.67
14	3.03 4.21	3.70 4.89	4.11 5.32	4.41 5.63	4.64 5.88	4.83 6.08	4.99 6.26	5.13 6.41	5.25 6.54
15	3.01 4.17	3.67 4.84	4.08 5.25	4.37 5.56	4.59 5.80	4.78 5.99	4.94 6.16	5.08 6.31	5.20 6.44
16	3.00 4.13	3.65 4.79	4.05 5.19	4.33 5.49	4.56 5.72	4.74 5.92	4.90 6.08	5.03 6.22	5.15 6.35
17	2.98 4.10	3.63 4.74	4.02 5.14	4.30 5.43	4.52 5.66	4.70 5.85	4.86 6.01	4.99 6.15	5.11 6.27
18	2.97 4.07	3.61 4.70	4.00 5.09	4.28 5.38	4.49 5.60	4.67 5.79	4.82 5.94	4.96 6.08	5.07 6.20
19	2.96 4.05	3.59 4.67	3.98 5.07	4.25 5.33	4.47 5.55	4.65 5.73	4.79 5.89	4.92 6.02	5.04 6.14
20	2.95 4.02	3.58 4.64	3.96 5.05	4.23 5.29	4.45 5.51	4.62 5.69	4.77 5.84	4.90 6.02	5.01 6.09

Q table: The critical values for q corresponding to alpha = .05 (top) and alpha = .01 (bottom)

When you look at this you see 4 is c, 20 is your n - c so that corresponding value when alpha equal 0.05 and 3.96 okay the Q table the critical values for Q corresponding to alpha equal to 0.05 on top and 0.01 at the bottom. This is the ANOVA table this on our table says you see that MSE is 6.51 mean squared error MS treatment is 127.6 because the value of 6.51 will use in the next slide.

(Refer Slide Time: 16:00)

The Tukey-Kramer Procedure

3. Compute Critical Range:

$$\text{Critical Range} = Q_U \sqrt{\frac{\text{MSW}}{2} \left(\frac{1}{n_i} + \frac{1}{n_j} \right)} = 3.96 \sqrt{\frac{6.51}{2} \left(\frac{1}{6} + \frac{1}{6} \right)} = 4.124$$

4. Compare:

$$\begin{aligned} |\bar{x}_1 - \bar{x}_2| &= |10.00 - 15.67| = 5.67 \\ |\bar{x}_1 - \bar{x}_3| &= |10.00 - 17.00| = 7 \\ |\bar{x}_2 - \bar{x}_3| &= |15.67 - 17.00| = 1.33 \\ |\bar{x}_1 - \bar{x}_4| &= |10.00 - 21.17| = 11.17 \\ |\bar{x}_2 - \bar{x}_4| &= |15.67 - 21.17| = 5.5 \\ |\bar{x}_3 - \bar{x}_4| &= |17.00 - 21.17| = 4.17 \end{aligned}$$

So, third step is compute the critical range we will find the critical train QU which you got from the table root of MSW is 3.651 one which I shown in the previous table 5.6 divided by 2 1 by 6 + 1 by 6 because same sample size, so that value is 4.124 then we will find the difference of two pair for example x1 and x2 the difference is 5.67 absolute difference is 5.67 but that is greater than 4.12 so we can say $\mu_1 \neq \mu_2$.

But look at this x2 and x3 this is less than your 4.124 so we will say $\mu_2 = \mu_3$ this is the observation which you got previously also so the mean of population 2 and population 3 is same, there is μ_2 and μ_3 same.

(Refer Slide Time: 27:00)

The Tukey-Kramer Procedure

5. Other than $|\bar{x}_2 - \bar{x}_3|$, all of the absolute **mean differences are greater than critical range**. Therefore there is significant difference between each pair of means, except 10% concentration and 15% concentration at the 5% level of significance.

Other than X 2 bar - X 3 bar when absolute value all of the absolute mean differences are greater than critical range therefore there is a significant difference between each pair of the means except 10 and 15% of concentration at 5% significance level. So, only these two concentrations there is no difference in tensile strength all other pairs are different.

(Refer Slide Time: 27:25)

Jupyter code

```
In [53]: from statsmodels.stats.multicomp import pairwise_tukeyhsd
from statsmodels.stats.multicomp import MultiComparison
mc = MultiComparison(data_r1['value'], data_r1['treatments'])
mresult = mc.tukeyhsd(0.05)
mresult.summary()
```

Out[53]: Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	lower	upper	reject
concentration10	concentration15	1.3333	-2.7894	5.4561	False
concentration10	concentration20	5.5	1.3773	9.6227	True
concentration10	concentration5	-5.6667	-9.7894	-1.5439	True
concentration15	concentration20	4.1667	0.0439	8.2894	True
concentration15	concentration5	-7.0	-11.1227	-2.8773	True
concentration20	concentration5	-11.1667	-15.2894	-7.0439	True

This we cannot solve with the help of Python so we import from stats model dot stat stat multi comparison import pairwise underscore tukey honestly significant difference does that HSD from stats model dot stat stat multi comparison import multi comparison mc multi comparison data underscore r1 is a value which you have already you remember we have done in the previous lecture data underscore onwards the treatment.

So mc result equal to mc – tukey hsd for 0.05 alpha mc result dot summary we will get this result. So, students what you have to do I have taken the screenshot of the output you have to enter this command into the Python command prompt then able to enter and verify the result. So, what is happening here is a group 1 group 2 you see 10%, 15% here false the rejection that means the rejection is false only that means these when mu equal to 10% of concentration and when the mu equal to 15% of concentration the means are equal all other pairs means are not equal.

(Refer Slide Time: 28:48)

Problem 2

- Following table shows observed tensile strength (lb/in square) of different clothes having different weight percentage of cotton.
- Check whether having different weight percentage of cotton, plays any role in tensile strength (lb/in square) of clothes.



We will do another problem the following table shows observed tensile strength found in square of different clothes having different weight % of cotton check whether having different weight percentage of cotton plays any role in tensile strength what we are going to do in this problem whenever the percentage of cotton is added into thee into the clothes this tensile strength is increasing. We will see that is there any connection between percentage of cotton and the tensile strength of the clothes.

(Refer Slide Time: 29:23)

Problem 2

Weight Percentage of cotton	Observed tensile strength (lb/in square)					Total	Average
	1	2	3	4	5		
15	7	7	15	11	9	49	9.8
20	12	17	12	18	18	77	15.4
25	14	18	18	19	19	88	17.6
30	19	25	22	19	23	108	21.6
35	7	10	11	15	11	54	10.8
						Grand total=376	Grand mean= 15.004



So here the weight percentage is taken in drove 15% 20% 25 30 35 this was the 5 observationally is given total is given the grand total is 376 the grand mean is 15.07.

(Refer Slide Time: 29:39)

- $SSA = 5(9.8 - 15.04)^2 + 5(15.4 - 15.04)^2 + 5(17.6 - 15.04)^2 + 5(21.6 - 15.04)^2 + 5(10.8 - 15.04)^2 = 475.76$

$SST = 636.96$

$SSE = 636.96 - 475.76 = 161.20$

Sources of variation	Sum of squares	Degrees of freedom	Mean square	F-value
Cotton weight percentage	475.76	4	118.94	14.76
Error	161.20	20	8.06	
Total	636.96	24		

First we will find out SS treatment some books they follow in SSB that means between sum of square SSA among sum of square some book write SS treatment, treatment sum of square. So, for treatment sum of square see in the row 1 there is a in the treatment 1 there are 5 element is there so 5 into this is the mean of first row. This is overall mean so 5 into 9.8 15.04 square plus this is their 5 element and 15.4 is the mean of the second row this is the overall mean Plus this is the mean of third row this is mean of third row mean of 4th row mean of your third row.

It is very getting SST treatment sum of square among the column that is otherwise SS treatment sum of sum of square is 475.76 similarly we have done previously with the help of our problems we can find out SSE we know this SST is 636.96 so if you want to know SSE 636.96 so we are getting 161.60 so this is ANOVA setup so this is sum of square of cotton weight percentage sum of square of error there is a degrees of freedom because there is a 5 rows, so 5 - 1 4 rows so there are 25 elements 25 - 1, 24 so 24 - 4 is 20 degrees of freedom.

When you divide by this 475.76 by 4 you are getting 118.94 when you divide 161.2 divided by 20 we are getting 8.06 so the F value is 4.76.

(Refer Slide Time: 31:19)

Problem 2

- When $\alpha = .05$, $F_{(0.05,4,20)} = 2.87$
- Reject H_0

```
In [17]: scipy.stats.f.ppf(1-0.05, dfn=4, dfd=20)
```

```
Out[17]: 2.8660814020156584
```

When alpha equal to because we could 0.95 because 5% means 0.95 numerator degrees of freedom is 4 degrees of when we are getting so this value when alpha equal to 5% this is 2.8 okay but our calculated F value is 14.76 so 14.76 will be this side which is on the right hand side so we have to reject our null hypothesis. After rejecting null hypothesis we refer Q table.

(Refer Slide Time: 31:59)

Problem 2

$$T_{\alpha} = q_{\alpha}(c, n - c) \sqrt{\frac{M S_E}{n}}$$

$$\alpha = 0.05$$

$$q_{0.05}(5, 20) = 4.23$$

$$T_{0.05} = 4.23 \sqrt{\frac{8.06}{5}} = 5.37$$

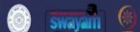
So, the Q value which we got from the table is when alpha equal to 5% is in his 4.23 so when it is a 4.23 MSC is 8.06 we got 8.06 from this one this value MSC 8.06 this, this value is taken as 8.06 turned by n so it is a 5.37. So, this 5.37 you have to compare any pair of treatment averages that differ in absolute value by more than 5.37 would imply that corresponding pair of population means are significantly different.

(Refer Slide Time: 32:37)

Problem 2

$$\bar{y}_1 - \bar{y}_2 = |9.8 - 15.4| = 5.6^*$$
$$\bar{y}_1 - \bar{y}_3 = |9.8 - 17.6| = 7.8^*$$
$$\bar{y}_1 - \bar{y}_4 = |9.8 - 21.6| = 11.8^*$$
$$\bar{y}_1 - \bar{y}_5 = |9.8 - 10.8| = 1$$

Starred values indicate pairs of means that are significantly different.

$$\bar{y}_2 - \bar{y}_3 = |15.4 - 17.6| = 2.2$$
$$\bar{y}_2 - \bar{y}_4 = |15.4 - 21.6| = 6.2^*$$
$$\bar{y}_2 - \bar{y}_5 = |15.4 - 10.8| = 4.6$$
$$\bar{y}_3 - \bar{y}_4 = |17.6 - 21.6| = 4$$
$$\bar{y}_3 - \bar{y}_5 = |17.6 - 10.8| = 6.8^*$$
$$\bar{y}_4 - \bar{y}_5 = |21.6 - 10.8| = 10.8^*$$


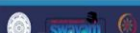
So, we have compared all comparison y_1 versus going to y_1 y_3 what is happening this is more than 5.37 this is this is here we can say $\mu_1 \neq \mu_2$ here also $\mu_1 \neq \mu_3$ but what is happening here it is less than less than 5.37. So, here we can say $\mu_1 = \mu_5$ here also $\mu_2 = \mu_3$ here also $\mu_2 = \mu_5$, here also $\mu_3 = \mu_4$ okay this is the Tukey Kramer test.

(Refer Slide Time: 33:22)

Jupyter code

```
In [2]: df3 = pd.read_excel('C:/Users/Somi/Documents/cotton weight.xlsx')

In [12]: data1 = pd.melt(df3.reset_index(), id_vars=['index'], value_vars=['cotwt.15', 'cotwt.20', 'cotwt.25', 'cotwt.30', 'cotwt.35'])
        data1.columns = ['id', 'treatments', 'value']
```



Then we will do with the help of Python I have imported the data calling it `df3 = pd.read_excel('C:/Users/Somi/Documents/cotton weight.xlsx')` underscore Excel because I have saved in the excel format then I am using this melt command you know that previously in the two classes I have used how to use this melt what is application

of this pdf dot so period to melt DF 3.2 reset underscore index id underscore versus in square bracket index value underscore versus cotton percentage 15 cotwt 20 cotwt 25 30 35 so date data 1 dot columns equal to id, treatment, values.

(Refer Slide Time: 34:08)

Jupyter Code

```
In [16]: mc = MultiComparison(data['value'], data['treatments'])
mcrests = mc.tukeyhsd(0.05)
mcrests.summary()
```

Out[16]: Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	lower	upper	reject
cotwt.15	cotwt.20	5.6	0.2266	10.9734	True
cotwt.15	cotwt.25	7.8	2.4266	13.1734	True
cotwt.15	cotwt.30	11.8	6.4266	17.1734	True
cotwt.15	cotwt.35	1.0	-4.3734	6.3734	False
cotwt.20	cotwt.25	2.2	-3.1734	7.5734	False
cotwt.20	cotwt.30	6.2	0.8266	11.5734	True
cotwt.20	cotwt.35	-4.6	-9.9734	0.7734	False
cotwt.25	cotwt.30	4.0	-1.3734	9.3734	False
cotwt.25	cotwt.35	-6.8	-12.1734	-1.4266	True
cotwt.30	cotwt.35	-10.8	-16.1734	-5.4266	True

So, MC is multi comparison data one value, data treatment one so mc result is equal to mc dot tukey dot hsd 0.05 when you MCE result dot summary when you type this what is happening here this, this, this that means the corresponding means are equal. So, other places not equal, so we have got whatever we got that result we have checked with the help of Python also. Dear students in this lecture what you have seen we have solved one way ANOVA problem in that one way ANOVA we have rejected null hypothesis.

Once we reject null hypothesis we have to say which two pairs of means is equal or not equal for that we have gone for and another set of test called post hoc analysis there are two test was there one is the least significant difference method another rule is Tukey Kramer method and also this be assaulted another problem with the help of Python, thank you.