

**INDIAN INSTITUTE OF TECHNOLOGY ROORKEE**

**NPTEL**

**NPTEL ONLINE CERTIFICATION COURSE**

**Marketing Research**

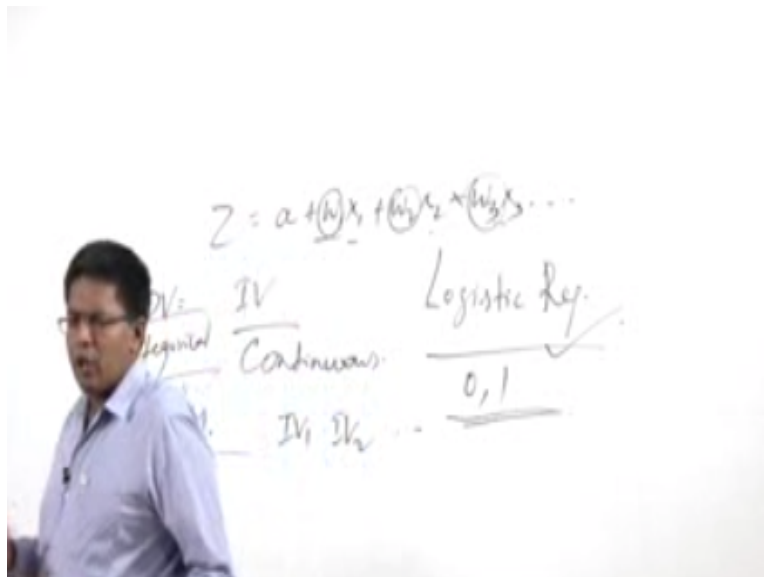
**Lec -33**

**Discriminate Analysis**

**Dr. Jogendra Kumar Nayak**  
**Department of Management Students**  
**Indian Institute of Technology Roorkee**

Welcome everyone to the section discriminant Analysis in this last section we had just started with an overview of what is discriminant analysis and what is the utility of discriminant analysis why is it used right and there we said that there are 2 basic techniques which are used for the similar purpose of what discriminate analysis does so what is discriminant analysis, discriminate analysis basically discriminates between 2 or 3 groups right multiple groups and gives us score that helps us to separate this groups in a proper way right.

(Refer Slide Time: 01:06)



So let us say there are several cases where are dependent variable right is let us say in categorical scale right has been written in a categorical way and the independent variables are rather in are

there in a continuous way continuous okay so when this condition occurs to us so that is the right time when we use discriminant analysis or there is another technique which is also used for the same called logistic regression.

A special case of regression logistic regression okay so the only difference between the logistic regression and the discriminant analysis is that the discriminant analysis is although it gives a it takes multiple you know categorical values for example let us say there are 3 groups 1, 2, 3, it can take that and it can even you know 3, 4, it can give a picture of how this groups are different from each other.

But on the other hand the logistic is generally applied for only cases where the 2 possibilities 0 and 1 yes or no kind of right so will happen or will not kind of a case but the advantage of a logistic regression is that a logistic regression does not get affected by the normality you know if there is violation of the normality of the data so violation of normality if it is still there logistic regression works well right but on the other hand if there is a normality violation discriminant analysis will have resource some problems okay.

But otherwise both the techniques are equally strong and power and widely used right and so this is what it does do it has helped you it helps us to find to discriminate between 2 or 3 groups right so we can say with the help of the independent variables with the help of independent variable whatever coefficients or weights you get by including that you can create a score which we say that discriminant score the score.

Now that score helps us to you know correctly say okay weather a new candidate or a new case that will come into the picture will fall into which group for example let us say there are 4 groups right so these groups there is a deviation on bases of certain parameter this parameters are the independent variables let us say so independent variable 1 independent variable 2 and goes on, so it taking this values from the past suppose let us say as I said earlier there is a bank the bank as given as given lone to people and some people have default the lone some people have given the lone in et right time.

They have you know before they have submitted their EMI money so by taking these 2 people kind of people you know a discriminate score can be calculated which is very similar to a regression only right so it goes on right so where  $x_1, x_2, x_3$  are your independent variables so by

taking this then the weights are the coefficients by including those F coefficients what it does is it helps you to suppose you get a new client or a new person who has come approach the bank for a loan then by placing his values the  $x_1, x_2, x_3$ .

And the coefficients it can using the coefficients it can say okay whether the new person would be a defaulter or a very a productive you know person for the bank so that is a basic utility so let us take an example so discriminant has his very popular and highly utilized in all kinds of financial markets you know even in decisions have to be taken either case of a like categorical way whether I will do it I will not do it you know kind of a thing so this is the case which I have brought in example, now this case what it does is basically it explains that two types of skull.

Now skull this is you know very interesting case what does happened is there are two types of skulls one found from the area of scheme and the you know close by area of tipper zone right and another kind was the you see type A and type B skulls right.

(Refer Slide Time: 05:58)

## Discriminant Analysis

The data can be divided into two groups. The first comprises skulls 1 to 17 found in graves in Sikkim and the neighbouring area of Tibet (Type A skulls). The remaining 15 skulls (Type B skulls) were picked up on a battlefield in the Lhasa district and are believed to be those of native soldiers from the eastern province of Khams.

These skulls were of particular interest since it was thought at the time that Tibetans from Khams might be survivors of a particular human type, unrelated to the Mongolian and Indian types that surrounded them.

The other one was from the Lhasa district right another place right so we are not interested in the place, so we are only using it has type A type B skull so now the, the researcher wants to know right it has got 5 criteria.

(Refer Slide Time: 06:14)

# Discriminant Analysis

The data examined here consist of five measurements on each of 32 skulls found in the southwestern and eastern districts of Tibet.

1. Greatest length of skull (measure 1)
2. Greatest horizontal breadth of skull (measure 2)
3. Height of skull (measure 3)
4. Upper face length (measure 4)
5. Face breadth between outermost points of cheekbones (measure 5)

There are also location and grouping variables.

Now this 5 criteria are it is 5 independent variables so what are the independent variables the greatest length of the skull right the greatest horizontal breadth of the skull that means what is the weight of the skull the height of the skull okay, the upper phase length then the cheat bone distance okay so this 5 you know variables have been used and a equation has been build okay with the help of the 32 skulls that where available to the researcher right, now he hopes that in that in the future.

Suppose a new skull is found he could be able to discriminate and stay whether this skull on basis of it is you know the measurements is it from there are the Sikkim zone Sikkim region or is from the Lhasa region okay, so it is a way of discriminating only correct so let us see how to do that so what are the two the two questions that might you have interest.

(Refer Slide Time: 07:23)

## Discriminant Analysis

There are two questions that might be of interest for these data:

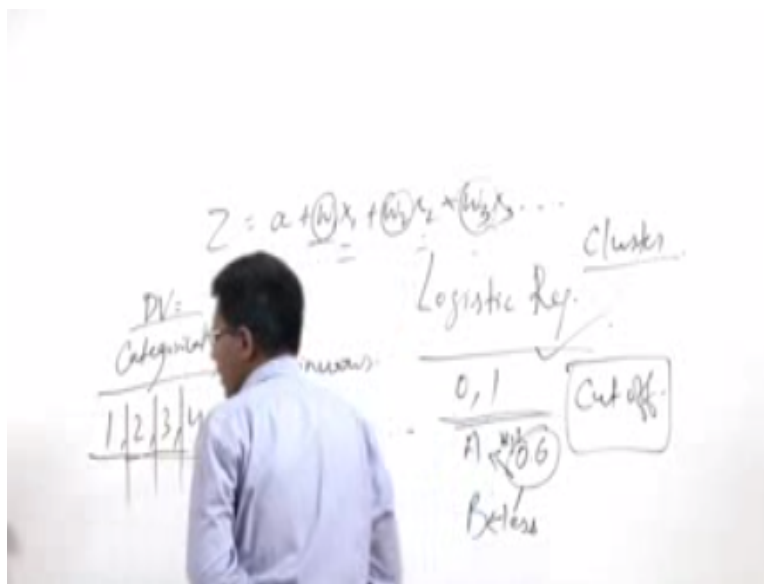
Do the five measurements discriminate between the two assumed groups of skulls and can they be used to produce a useful rule for classifying other skulls that might become available?

Taking the 32 skulls together, are there any natural groupings in the data and, if so, do they correspond to the groups assumed?

Are do the five measurements discriminate between the two assumed groups of skulls and can they be used to produce the useful rule for classifying the other skulls right that might become available okay, so taking the 32 skulls right they have try to form a as I said a group right and this two groups type A and type B and a discriminate score is to be built and on basis of that we will be able to decide whether it is which kind of skull okay so as I said as it is said if you can see discriminant analysis helps in classification but when I say classification you might be remembering the in the last to last somewhere.

We had discussed about cluster analysis cluster analysis also an important tool to classify if I that time I had said.

(Refer Slide Time: 08:20)



The cluster is basically used to classify so then what is the difference between this cluster and discriminant cluster only creates groups when it has not there right so cluster uncovers groups it develops the groups.

(Refer Slide Time: 08:35)

## Discriminant Analysis

Classification is an important component of virtually all scientific research. Statistical techniques concerned with classification are essentially of two types.

The first (cluster analysis) aims to uncover groups of observations from initially unclassified data.

The second (**discriminant analysis**) works with data that is already classified into groups to derive rules for classifying new (and as yet unclassified) individuals on the basis of their observed variable values.

It develops the groups right from observations of initially unclassified data okay on the other hand the discriminant analysis derived rules on basis of the already classified data right, and now clubs the individual re places the individual on to the right group according its features, features are the independent variables right, so that is the basic different so how do you do it I have just brought an example with through how it you know to do it through SPS there few values which are very important when you talk about discriminant.

Now discriminant one thing is there is something called a discriminant score or a you know or cut off score also right cutoff score now once you have let say created a data and you have decided a cut of value let say the cutoff value is let say somewhere let say 0.6 right now the cutoff value is 0.6 then you say there are two groups and on basis of this cutoff value we will say okay whether the by or the person will fall into which category which are okay group 1 or group A or group B A or B if it is let say it is greater than 0.6 it is A let say if it is less than then suppose group B if it **is** high then group A.

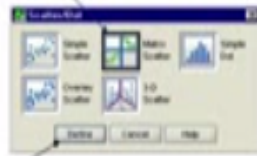
So to find the cut off value is also equally very important for the researcher. Now how does he find the cut off value we will see that, okay so first of all let us go through a scatter plot just we will decide scatter plot of the data, right.

(Refer Slide Time: 10:25)



# Discriminant Analysis

Select matrix scatter



Use Define to select.


(Refer Slide Time: 10:26)

## Discriminant Analysis

Select matrix variables and markers.

Note that greatest length of skull is above the list shown.

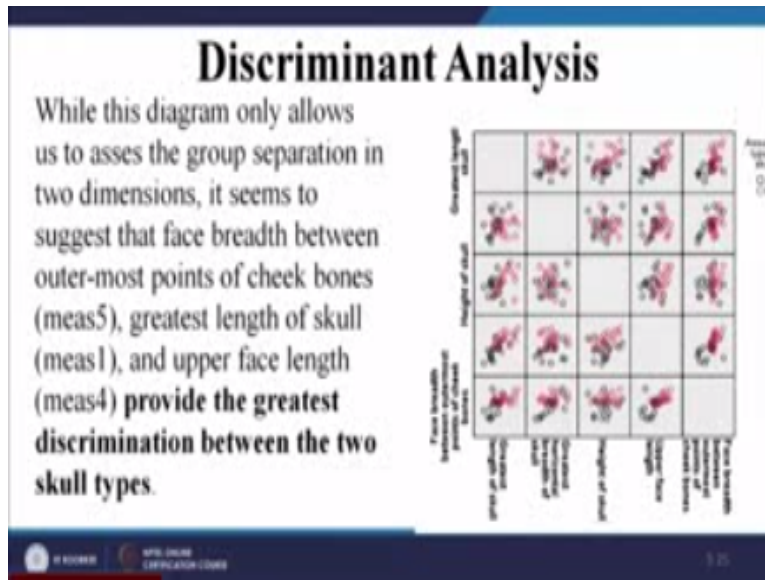
Use OK to accept, or Paste to preserve the syntax.



The screenshot shows the SPSS Discriminant Analysis dialog box. On the left, there is a list of matrix variables: 'Greatest horizontal breadth', 'Height of skull (mm)', 'Upper width length', and 'Phase width'. The 'Classify by' section is set to 'Assumed type of skull (mm)'. The 'Display' section has 'Display discriminant function coefficients' checked. The 'Paste' section has 'Paste discriminant function coefficients' checked. The 'OK' button is highlighted.

Taking all the matrix variables you know the horizontal breadth, height, upper width length, phase width and said the markers on basis of the assumed type of skull type A or type B so the researcher knows okay, let us say skull number 3 let us say is from seek him right and the features where these okay, skull number 7 was from let us say lasa and these were the features, so he knows so he has set markers by the type of skull okay.

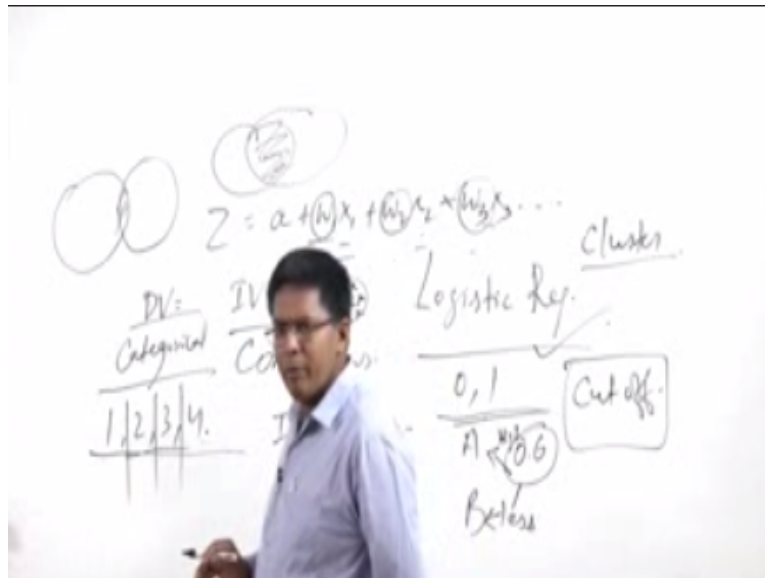
(Refer Slide Time: 10:59)



Now if you look at this diagram although generally it is not very clear one cannot be very specific from a diagrammatic representation at least I believe that right, so what you do is if you look at this it allows us to assess the group separation in two dimensions right, group A and B it seems to suggest that face breadth it says from between the out most points greatest length of the skull and the 154, 1 it is like a correlation right, so correlation.

So 154, 1 this is the chick bone distance is this one, 5 then you have 1 and then you have 4 these three if you see let us look at them and if you see the data the you know the type of skulls right, their clearly there is separation but in other cases in this case for example it is super imposed so when the most important thing is discriminate analysis is that.

(Refer Slide Time: 12:09)



Suppose there is a too much of there is a little overlap then we will say that it discriminates well, but what if the data is highly you know super imposed so if this is the area then it is too much of super imposition in this case we cannot say that it is discriminating well, right. Similarly when data in suppose this is you know like the data is there and suppose another kind of data is let us say this is what it is showing right, suppose it is interest imposed in between so then you cannot discriminate there has to be a clear pattern of discrimination. So let us see how it is being done, okay.

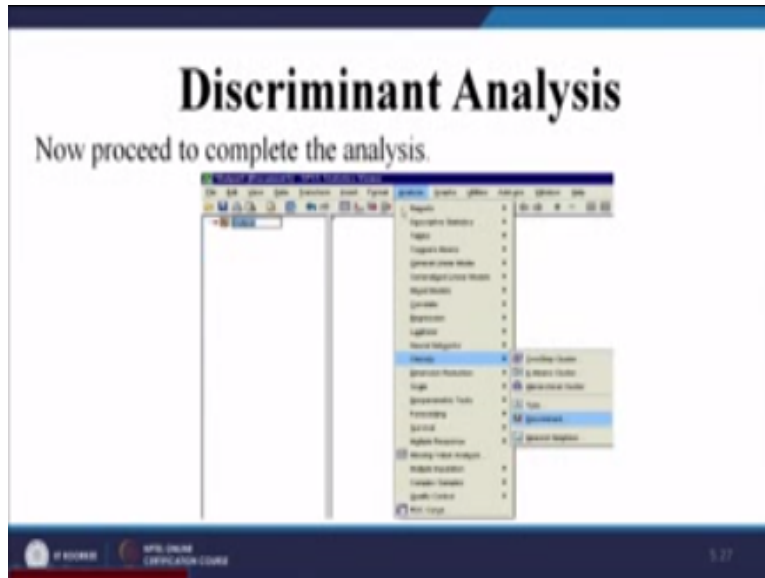
(Refer Slide Time: 12:51)

## Discriminant Analysis

We shall now use Fisher's linear discriminant function to derive a classification rule for assigning skulls to one of the two predefined groups on the basis of the five measurements available.

So we use the Fisher's linear discriminate function, so Fisher's linear discriminate function is I think this is what is the one the Z right, so we use this Z value and we are going to measure this Z value right, so with the help of this Z value once you have the Z value then it becomes easier for you right, to know okay which group it comes to okay, let us see. Now how do you do that, so in the SPSS for example I am using SPSS.

(Refer Slide Time: 13:24)

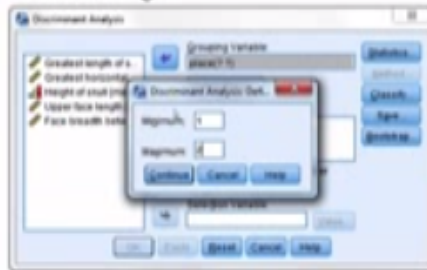


So if you go to a classify analyze then classify and then there is discriminant right.

(Refer Slide Time: 13:31)

# Discriminant Analysis

As before use the secondary screens to select the grouping variable (place) and use Define Range.



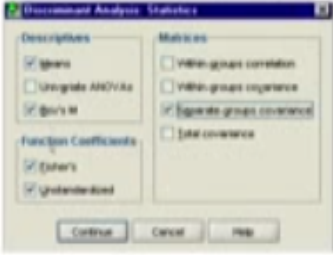
So once you go there you can say what the minimum number of you know your grouping variable, so in case let us say there are two groups to be formed only Sikkim and Lasa so to two groups right, so 1 and 2 but suppose there are more than two groups there could be three groups also, so the highest value would be 3 the lowest value would be 1 right, so that would differ suppose it is 4 then 11 and 4.

But remember you should not be taking too many you know your groups also because you can understand why to I am saying, if you take too many groups the discriminating power will be very complex and will not be very clear it will in one way reduce okay.

(Refer Slide Time: 14:17)

# Discriminant Analysis

From the statistics button make the following selection



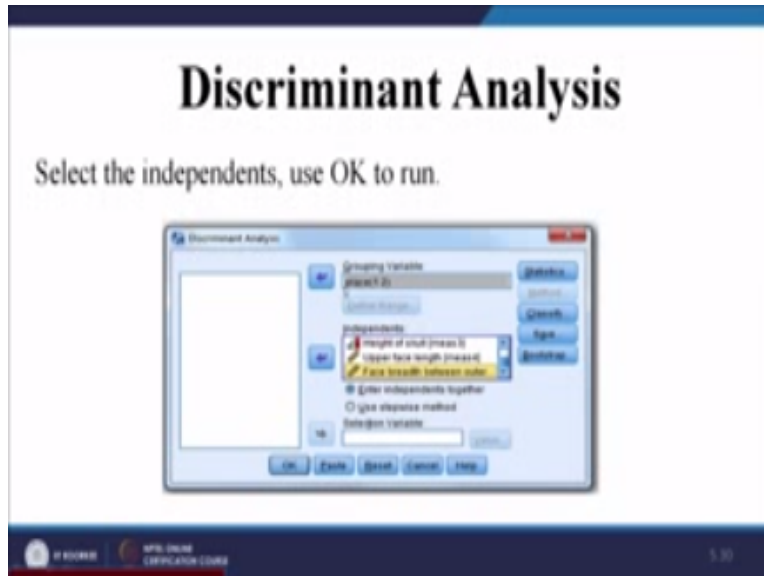
Now proceed to complete the analysis.

SPSS 2019  
SPSS ONLINE  
CERTIFICATION COURSE

So now we have taken the means this is the Fisher's function coefficient and the un-standardized there is a difference between the standardized and the un-standardized, the standardized is good because in some cases the standardized helps us to compare but un-standardized is helpful also because it fits in the raw data and gives you a very, it is very easy to calculate right that is the basic difference.

(Refer Slide Time: 14:41)





So once we have taken this grouping variable and the independent variables there are two methods, so either we put in all the variables together simultaneously so when you take simultaneously the obviously the system will take it together and then you know give you a coefficient. But in there is another way also where if you can use like regression you can use it in a step wise method so step method what it does is basically in step wise method it calculates it how does it select a variable it selects the variable on basis of the distance the mahalanobis distance right so the mahalanobis distance right,  $d^2$  we say  $d^2$ .

So mahalanobis distance is the internal the algorithm is on that way right the higher the suppose there are different value let us say 0.3 0.7 1.4 1.2 0.8 now what it does is it tries to pick up the one with the highest value right and first feeds into the system right and then it goes on right.

(Refer Slide Time: 16:04)

# Discriminant Analysis

The within-group covariance matrices shown in the Covariance Matrices table suggest that the sample values differ to some extent, see Box's test for equality of covariances (see Log Determinants and Test Results, below).

Covariance Matrices

Face shape		Forehead height of skull	Orbital breadth of skull	Height of skull	Upper face width	Face breadth between external points of outer ear
Sikkim (N=12)	Forehead height of skull	22.222				
	Orbital breadth of skull	28.222	37.889			
	Height of skull	12.222	11.889	38.222		
	Upper face width	22.222	7.889	-2.222	28.222	
	Face breadth between external points of outer ear	17.889	48.889	1.889	18.222	88.222
	Mean	19.444	8.556	22.778	17.778	11.111
Lasa (N=12)	Forehead height of skull	8.333	37.889			
	Orbital breadth of skull	22.778	11.333	38.333		
	Height of skull	17.778	7.889	18.333		
	Upper face width	11.333	7.889	18.333	18.333	
	Face breadth between external points of outer ear	11.333	8.889	7.889	8.889	17.889
	Mean	11.111	19.444	18.333	18.333	11.111

So if you look at the group statistics now these are the descriptive variables the mean and the standard deviation right for the Sikkim people and the Lasa people now this will give you an idea what is the actual difference between the Sikkim the Lasa people, but it does not say anything beyond this it does not talk about the significance of the difference right it does not do that.

So here our interest is to not to look in to this things but rather we want to see is there any way or is there you know out of this five variables which variable or which variables are the one's which are able to discriminate between the two groups more powerfully, now for example suppose we had five we said right so higher the let us say w higher the w value generally that will have the larger impact okay.

So within group co variants now when I talk about co variants matrices okay let us see this, so what we have basically doing it is giving you a log determinants in the next slide right so this is the values greatest lambda scale you know greatest all these are given to you and in the table which is more important.

(Refer Slide Time: 17:25)

# Discriminant Analysis

The within-group covariance matrices shown in the Covariance Matrices table suggest that the sample values differ to some extent, but according to Box's test for equality of covariances (tables Log Determinants and Test Results) these differences are not statistically significant ( $F(15,3490) = 1.2, p = 0.25$ ).

Place where study was done	Rank	Log Determinant
Abroad in Year	3	18.786
Home	3	18.779
Worked within group	3	18.727

The rank and value regardless of determinants listed are those of the group covariance matrices.

Statistic	Value
F	1.211
df1	15
df2	3490
Sig.	.249

Tests for homogeneity of mean population covariance matrices.

Now because earlier table is not that requirement to us so this two tables now log determinants means it has been converted the ranks and the natural log of the determinants are given to you that means the original values have been converted in to a log value right and the test results are given so although this test results right what they are saying? They measure the equality of covariances this differences are not spastically significant, now how did they know? From this value from this value if you see 0.249 it suggest that because since you might be testing at 95% confidence level or 99% confidence level if it is more than the significance level then the  $\alpha$  then that case you reject you except the null hypothesis.

So here our null hypothesis is that here is not difference so this differences are not statistically significant had they been less than 0.05 there is a statistical difference there is the significant difference between the values but it does not happen here it is not happening right. So now coming to this table.

(Refer Slide Time: 18:39)

# Discriminant Analysis

The resulting discriminant analysis shows the eigenvalue (here 0.93) represents the ratio of the between-group sums of squares to the within-group sum of squares of the discriminant scores. It is this criterion that is maximized in discriminant function analysis.

Eigenvalue

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	0.93	100.0	100.0	.694

\* First 1 canonical discriminant functions were used in the analysis.

Now this table if you see there are two important values given to you one is the Eigen value which it is explaining here is 0.93 right and the canonical co relation now what is this Eigen value the Eigen value in this case in a the case of discriminate analysis is nothing but the ratio of within sorry the ratio of not within the ratio of between by within okay the ratio of between by within between groups some of squares to within some of squares right.

So if you remember I had also explained you in a way to remember that within is generally we talk about the error it consist of the error right. So the explain this is the explained this is the unexplained okay. So this value the higher this Eigen value is the larger is the you know explanation of the variables, so the canonical what it says here is it is the criteria that is the criteria the maximize in the discriminate function analysis so if it is 0.93 that means it is there is a possibility that it is there are some the variables are clearly discriminating between the two group 1 and group 2 let say and what is the correlation.

If you remember just like we had  $r^2$  in regression we use to have  $r^2$  but that was the case of regression here we cannot use that because here our variables are not metric in nature they are not in metric form they are one in dependence in categorical and the independent variables in metric.

So here that is why we use the canonical correlation but the understanding remains the same for example in this case .694 or .7 let say .7 is my canonical correlation you squared it and when you

squared it suppose it is .4 sorry, sorry .49 let say so 49% of the explanation is happening with the help of this independent variables okay this is what it says okay.

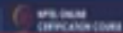

(Refer Slide Time: 21:07)

## Discriminant Analysis

The canonical correlation is simply the Pearson correlation between the discriminant function scores and group membership coded as 0 and 1. For the skull data, the canonical correlation value is 0.694 so that  $0.694^2 \times 100 = 48\%$  of the variance in the discriminant function scores can be explained by group differences.

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	828	100.0	100.0	.694

\* First 1 canonical discriminant functions were used in the analysis.

1/30

So if you look at this, this says that it has taken the exact score .694 so that si why it is coming 48% of the variance in the discriminate function scores can be explained by the group differences which is automatically impact of the measurement variables only right so that si what I was saying how the measurement variables are impacting it okay.

(Refer Slide Time: 21:36)

# Discriminant Analysis

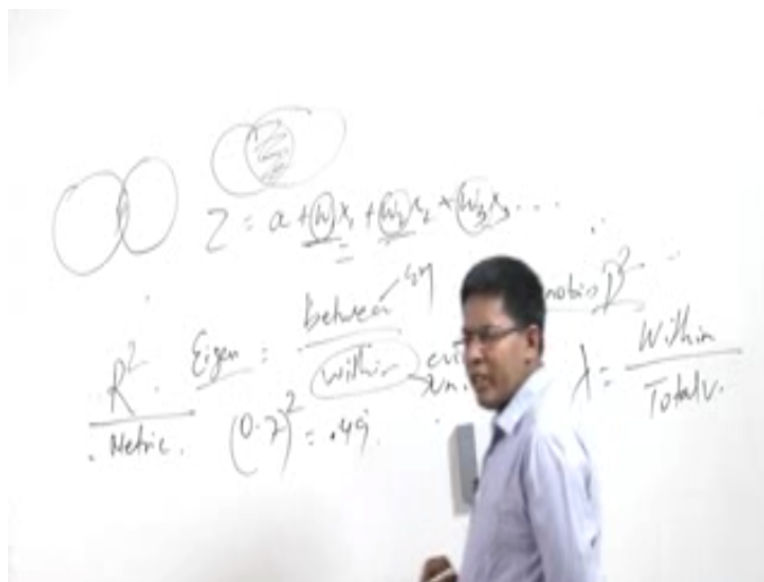
If Wilk's Lambda provides a test for assessing the null hypothesis that in the population the vectors of means of the five measurements are the same in the two groups. The lambda coefficient is defined as the proportion of the total variance in the discriminant scores not explained by differences among the groups, here 51.8%. The formal test confirms that the sets of five mean skull measurements differ significantly between the two sites ( $\chi^2(5) = 18.1, p = 0.003$ ).

WILK'S LAMBDA

Case of Function	Wilk's Lambda	Chi-Square	df	Sig.
	.482	18.101	5	.003

Now next is if you look at the something call Wilk's Lambda now this is very important again now Wilk's lambda is a process is used is the test provides a test for assessing the null hypothesis that in the population the means of the 5 measurements are the same in the two groups what is it say wilk's lambda says that there is no difference between the means of the 5 measurements right.

(Refer Slide Time: 22:18)



In both the groups that is second group and the last group and the lambda coefficient is defined as the proportion of the total variance you can understand this way the lambda the wilk's lambda is nothing but within variance divided by the total variances so within variance divided by total variance so if you that means what if my and I have told you that within something which is unexplained right.

So if I am having my lambda value of high lambda value so what does it indicate it indicates that there is no clear pattern or no clear discriminating power in the case in this case right so if you have the lower the lambda value the better is the discriminating you know the ability to discriminate the groups right.

So it says the lambda coefficient is .518 and it s significant right in this case so now if it is significant that mains what is the null hypothesis that the 5 measurements are the same right in the two groups so that is getting rejected that is getting rejected in this aces no there is no significant difference in the there is the significant difference in the two groups that means right.

(Refer Slide Time: 23:29)

## Discriminant Analysis

If Wilk's Lambda provides a test for assessing the null hypothesis that in the population the vectors of means of the five measurements are the same in the two groups. The lambda coefficient is defined as the proportion of the total variance in the discriminant scores not explained by differences among the groups, here 51.8%. The formal test confirms that the sets of five mean skull measurements differ significantly between the two sites ( $\chi^2(5) = 18.1, p = 0.003$ ).

WILK'S LAMBDA

Order of Functions	Wilk's Lambda	Chi-Square	df	Sig.
1	.482	18.111	5	.003

Because the null hypothesis was that they are the same so they are not same now okay now something come into the classification coefficient right now this classification function coefficient is generally used for you know you remember when you have a two groups the number of function coefficients you can develop is only one.

Because the rules says the function coefficients rea the values you have is always NG are number of groups -1 this is the formula so suppose you have a three group so then you have two function coefficients okay you can see this right this function canonical distribution functions coefficients you will get only one right so if you have more then accordingly it has to be NG-1 right.

(Refer Slide Time: 24:41)



# Discriminant Analysis

Next we come to the Classification Function Coefficients. This table is displayed as a result of checking Fisher's in the Statistics sub-dialogue box.

Classification Function Coefficients

	Place where skulls were found	
	Skim on Tibet	Lhasa
Greatest length of skull	1.408	1.538
Greatest horizontal breadth of skull	2.361	2.258
Height of skull	2.752	2.707
Lower face length	775	852
Face breadth between subnasal points of cheek bones	195	372
(Constant)	-214.858	-545.478

Fisher's linear discriminant functions

So now the question is you have brought the classification function coefficient now how can I use this now this is used to create our z values for example now x value in this case generally you do not this you know thing is not given in most of the it is not explained in a better way now how it is happen now this is how it has been done actually now if you see if you look at the values here of the of the skim and the Lhasa.

(Refer Slide Time: 25:07)

# Discriminant Analysis

It can be used to find Fisher's linear discriminant function as defined by simply subtracting the coefficients given for each variable in each group giving the following result:

	Skull or Total	Lower	Difference
Greatest length of skull (measure 1)	1.488	1.558	-0.099
Greatest horizontal breadth of skull (measure 2)	2.381	2.205	0.156
Height of skull (measure 3)	2.752	2.747	0.005
Upper face length (measure 4)	0.775	0.952	-0.177
Face breadth between extremal points of cheekbones (measure 5)	0.195	0.372	-0.177

$$Z = -0.09 \text{ meas1} + 0.156 \text{ meas2} + 0.005 \text{ meas3} - 0.177 \text{ meas4} - 0.177 \text{ meas5}$$

The difference between these two is used to calculate the Z score right, now the Z score for example you see  $Z = -0.09 \text{ measure 1}$  that is  $x_1 + 0.156 x_2, 0.005 x_3 - 0.177 x_4$  and similarly for  $x_5 - 0.177$ .

(Refer Slide Time: 25:33)

# Discriminant Analysis

$$Z = -0.09 \text{ meas1} + 0.156 \text{ meas2} + 0.005 \text{ meas3} - 0.177 \text{ meas4} - 0.177 \text{ meas5}$$

The difference between the constant coefficients (-514.956 and -545.419, bottom row of Classification Function Coefficients) provides the sample mean of the discriminant function scores

$$\bar{Z} = 30.463$$

Classification Function Coefficients

	Place where skulls were found	
	Open	Shaded
Minimum length of skull	1.208	1.538
Greatest horizontal breadth of skull	2.361	2.208
Height of skull	2.762	2.747
Upper face length	778	862
Face breadth between external points of cheek bones	188	372
Constant	-514.956	-545.419

Place's linear discriminant functions

No what and the difference between these two if you can see, these two values is nothing but the Z mean, the mean score right, it was the sample mean of the discriminative function scores right.

(Refer Slide Time: 25:51)

# Discriminant Analysis

The coefficients defining Fisher's linear discriminant function in the equation are proportional to the unstandardised coefficients given in the "Canonical Discriminant Function Coefficients" table which is produced when Unstandardised is checked in the Statistics sub-dialogue box.

Canonical Discriminant Function Coefficients

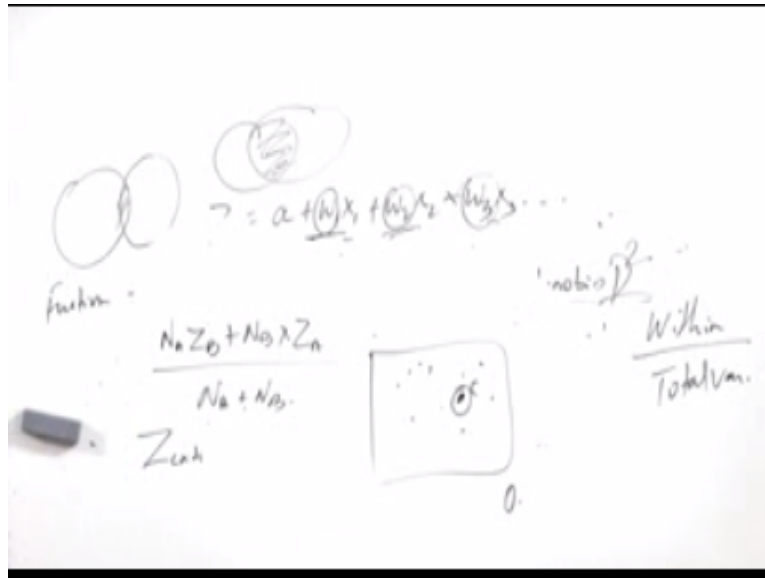
	Function 1
Standard length of fish	.000
Standard perimeter	.000
Length of snout	.000
Snout-vent length	.000
Eye length (mm)	.000
Standard points of head length	.000
Constant	-10.000

Unstandardized coefficients

Now I think let us go back and check right, so you have got this function coefficient, now there is something that is important to you now called a group centroids. Now what are the group centroids? The group centroids are nothing but when you have two groups in this case Sikkim the average of the means of the group scores, the group centroids is nothing but the average of the scores of the particular group.

For example let say whatever values you have, so if you plot those values right, it finds some mean values, which is this is the centroids, for example. Now this score this centroids becomes a very you know easier way of comparing right. Now for example in this case with the help of these centroids we find the cut off scores. Now the cutoff score which I earlier said, now what is this cutoff scores? Generally if you have two equal groups then the formula is  $Z(a) + Z(b)/2$ .

(Refer Slide Time: 27:11)



But in case your groups are unequal then this formula will change to  $N_a \times Z_b + N_b \times Z_a / N_a + N_b$ , so by this i can find the centroids right the cutoff value. The cutoff score whatever it comes, cutoff score is sometimes is also called as the Z critical right. Now that Z critical is compared with the Z value and accordingly if it is more or less you decide whether it will fall into this group or the other group.

For example in this case the function groups' centroids, for the Sikkim are 0.877, Lhasa is 0.994 right. now the centroids is been calculated the skulls sorry the cutoff scores has been calculate as 0.085, it has just taken the average, as we mean equal groups right. So the new skulls discriminate scores above 0.0585 because you have to understand. Now although there are two groups, the researcher as only one particular value with it right.

And this particular value that it has got, that has to be the way of comprising, 0.0585 would be assigned above anything 0.0585 it is assigned to the Lhasa which is 0.994. And if it is less than, if it is above 0.05 Lhasa, if it is less than 0.0585 it goes into the Sikkim right. So this value cutoff value is very important part or number in the discriminative analysis. Without this you cannot discriminate or you cannot place the new character, new personality or the new person or the new customer or the skull in this case. So these are some standardized value.

(Refer Slide Time: 29:41)

## Discriminant Analysis

For our data, such standardisation is not necessary since all skull measurements were in millimetres. Standardization should, however, not matter much since the within-group standard deviations were similar across different skull measures. According to the standardized coefficients, skull height (meas3) seems to contribute little to discriminating between the two types of skulls.

Standardized Canonical Discriminant Function Coefficients

	Canine
Skull height of skull	0.07
Cheek breadth	0.627
Height of skull	-0.17
Skull base length	0.51
Face breadth (maxilla)	0.55
Subnasal width of skull base	0.57

I am not getting into it right, so yeah this is one thing important, when you are doing a studies, suppose in this case for example all our measurements where using the same kind of you know scale, they are using millimeters right yes but there could be certain cases where you have used this discriminate analysis and the independent variables are using different kind of measurement scales so in that case it is necessary as a researcher to for you to understand that you need to standardized and use the standardized values if you do not use standardized values you are doing something completely you know big error right and now looking at the standardized discriminate function coefficient in this case we can see.

Okay out of this 5 variables the one which is contributing lowest to the discriminating power is 0.17 this is the weight they are the coefficient right is -0.17 this is the height of the skull so height of skull actually is not able to through height of skull you cannot say whether this skull is from skim or lacer but if you look at other one which is 0.627 phase breadth between the cheek bones so the distance between the cheek bones is very important parameter to decide whether the you know the skull is from skim or from lacer right so these again there I something -0.578 so this is also you need not worry about the direction rather you should be worrying about the you know the absolute score right.

(Refer Slide Time: 31:32)

## Discriminant Analysis

However, estimating misclassification rates in this way is known to be overly optimistic and several alternatives for estimating misclassification rates in discriminant analysis have been suggested.

One of the most commonly used of these alternatives is the so called leaving one out method, in which the discriminant function is first derived from only  $n - 1$  sample members, and then used to classify the observation left out. The procedure is repeated  $n$  times, each time omitting a different observation.

Now once you have done this.

(Refer Slide Time: 31:35)

## Discriminant Analysis

This is known as the re-substitution estimate and the corresponding results are shown in the Original part of the Classification Results table. According to this estimate, 81.3%  $((17 \times 82.4 + 15 \times 80) / (15 + 17))$  of skulls can be correctly classified as type A or type B on the basis of the discriminant rule.

	Percentage	Discriminant Group	Predicted Group		Total
			Skull A	Skull B	
Original	Total	Skull A	12	3	15
	Skull A	12	3	15	
	%	80.0	20.0	100.0	
Cross-validated	Total	Skull A	12	3	15
	Skull A	12	3	15	
	%	80.0	20.0	100.0	

1. Cross-validation is done only for those cases in the database. In cross-validation, each case is classified by the function derived from all cases other than that case.  
 2. 81.3% of original grouped cases correctly classified.  
 3. 80.0% of cross-validated grouped cases correctly classified.

This is something the classification matrix or the classes now once you have done how do you validate the question comes is I also must have said in the last session how do you validate to validate you know the results you have this one simple way that you can just buy for kid divide the entire data set into two parts one you use for analysis and other you use for let us say for as a hold out sample right and then you do the study on both and see whether there is any significant difference between the two or not right that is one thing.

So in this case when you have see when the original values are taken right so 14 skulls of skim actually fell into skim the group but by mistake 3 went to the laser group it is because of the confusion similarly laser there are three cases which actually fell into 12 fell into the laser group originally and three fell into skim group so now what is the accuracy result the accuracy result is so that means it is 81.3% right so this is how 17.82.4 now 82.4 is the percentage so 14/17 right and 15.80 right of laser so that gives me the classification power right similarly this is a cross validated result.

So in the cross validated result if suppose there is in this case there are significant decrease in think yes so in the cross validation is that 65.6% has been only explained so considerably lower success rate than suggested by the earlier one so when it happens then it becomes a very important point of thinking for the marker or the let us say the banker or anybody whether there is something else that we have missed or we need to find out why is the score let us say after the



cross validation why is the cross has come down from 81.3%to 65.6% right you can also measure it in the same way.

So this gives a feed back or it gives you a input to discuss or to decide what should the marketer do to in order to see that this accuracy rate does not differ to much right so this is all we have for this session I hope descriptive analysis is clear to you so you need to understand what is the discriminate score right like a recursion score you have descriptive score so and the like a coefficient here also you have a coefficient it is all same and it is the inverse of man ova which I have said it is a inverse of manova which I have said in the first class in the last session of descriptive analysis.

And finally you need to find out if there are two groups or three groups or four groups how do I decide which are the variables which are contributing highest to the explanation power and what is the cut off score to decide whether a particular candidate will fall into one group or the other right so this is what descriptive analysis does thank you very much for this session.

#### **For Further Details Contact**

**Coordinator, Educational Technology Cell  
Indian Institute of Technology Roorkee  
Roorkee 247 667**

**E-Mail [Etcellitrke@gmail.com](mailto:Etcellitrke@gmail.com) [etcell@itr.ernet.in](mailto:etcell@itr.ernet.in)**

**Website: [www.itr.ac.in/centers/ETC](http://www.itr.ac.in/centers/ETC). [www.nptel.ac.in](http://www.nptel.ac.in)**

#### **Production Team**

**Sarath Koovery**

**Mohan Raj. S**

**Jithin. K**

**Pankaj saini**

**Graphics**

**Binoy. V.P**

#### **Camera**

**Arun. S**

#### **Online Editing**

**Arun.S**

#### **Video Editing**

**Arun.S**

**NPTEL Cooridinator**

**Prof B.K Gandhi**

**An Educational Technology Cell**

**IIT Roorkee Production**

