

INDIAN INSTITUTE OF TECHNOLOGY ROORKEE

NPTEL

NPTEL ONLINE CERTIFICATION COURSE

Marketing Research

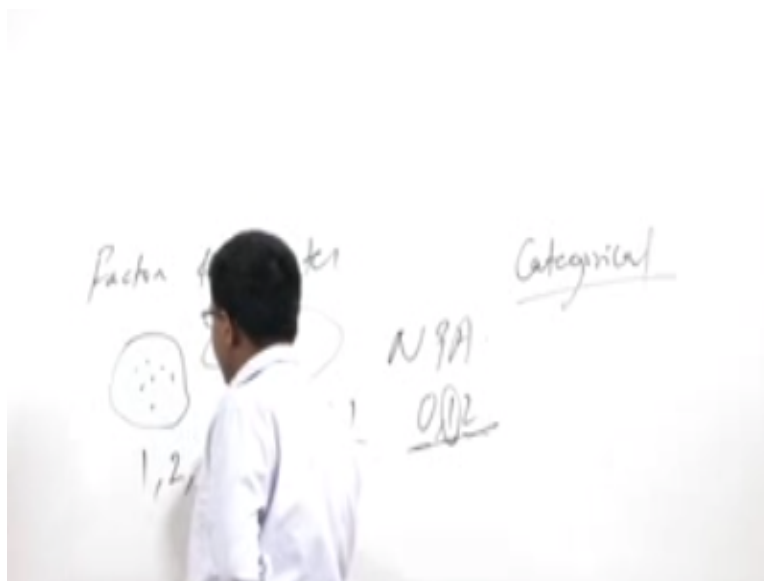
Lec -32

Discriminant Analysis

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

Welcome everyone to the class of marketing research and analysis in the last section we had discussed about a very important grouping method which was called the factor analysis and there after we discussed about another grouping method which was called the cluster analysis right so factor and cluster where two important techniques.

(Refer Slide Time: 00:41)



That where used to differentiate or to you know to create groups or clusters so that respondents could be categories or you know easily brought into different groups right clusters so that each cluster had it is own characteristic different from the other clusters okay so that was a very

important way of classification which we used to classify the cases the respondents the products anything you can say right.

But today we will talk about a situation where a marketer is faced with a very important observation let us say there is a bank which as to give a loan okay now the question is the bank as to decide whether the client is a worthy client or not right so in such a situation if it is not a worthy client and by chance the bank offers a loan to the client then the bank will lose its money right.

Maybe you know like many big corporate houses today are siphoning off the bank's money and not paying staying in London or somewhere so they are not paying the bank's money and the banks are completely getting into loss which has increased the non-performing assets of the banks, so it has become very critical that the bank should learn whether they should give a loan to this client or not similarly a credit card company wants to also find out whether they should give a credit card to the client or not right many other examples could be given for example whether should I take a candidate to my course or should I not take a candidate to my course right.

Some other cases where you want to know okay whether a particular person let us say falls into very let us say as per his religious aspirations into a highly religious moderately religious or not religious at all kind of a group right so whatever the condition is here if you see all your distributions are basically depending upon variables which are categorical in nature right so if my dependent variables are categorical in nature that means I can only say whether it is let us say for example the loan should be given yes or loan should not be given as 0 right.

So my diction is based on 0 and 1 right could be 0, 1, 2 in some cases right for example not given yes surely given moderate I do not know can be given or not given cannot say kind of right somebody is highly religious somebody is moderately religious somebody is let us say not religious however you want to break it up so the question emerges theories is this obviously I think you understood okay how important this is right so the researcher or the marketer wants to decide or separate the clients on basis of certain variables right, so the question is what how will you do this so do this such kind of a you know to take such kind of a decision we use to effect techniques one is called discriminant analysis.

(Refer Slide Time: 04:31)



Discriminant analysis the other is the form of regression only which is called logistic regression okay so logistic regression and discriminate analysis are two types of techniques which are used very largely so that the decision maker can take a decision which is based on a categorical level right or either or with the or the decision is like a categorical variable right, but his decision is based on certain independent variables which are all continuo in nature are all continuous in nature so as the result.

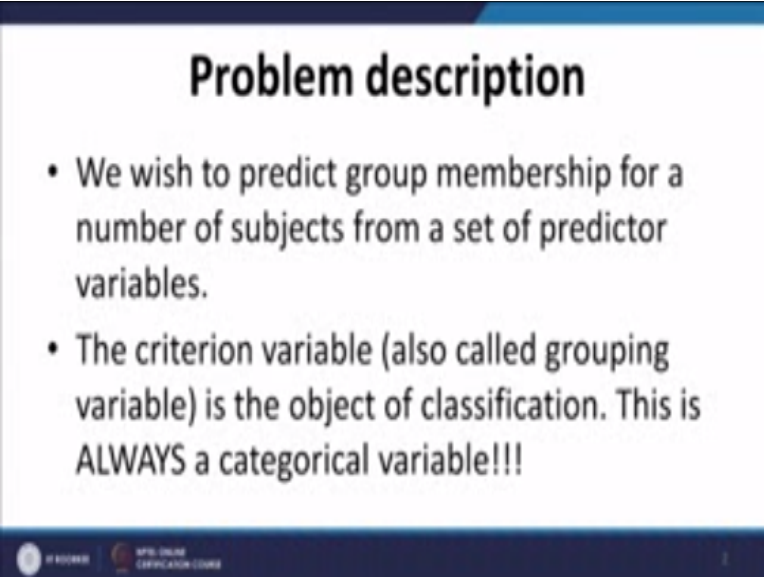
The researcher has to so the what is the way how the researcher is going to you know work can such a problem right do this is as I explained we do have discriminant analysis and logistics regression before I say if I go further what is the basic difference all though logistic regression is a special case of regression which I will be taking up in another maybe session, today I will be talking about the discriminate analysis the discriminant analysis and the logistic regression that difference although we today have multiple.

You know multi nominal logistic also but the advantage of discriminant is the discriminant is that discriminant is little more powerful than the logistic but on the other hand discriminant is also highly such to certain assumptions right for example the data should be has to understand okay discriminant analysis and logistic regression although they do the same thing same job, but discriminant is more little more powerful in compression to logistic because logistics is generally

done with 0 and 1 case yes or no case but discriminant analysis can take up more than 2 3 4 also.

A categories but larger to larger numbers of categories also is not very advisable okay, so let us see let us take a first case we wish to predict membership.

(Refer Slide Time: 06:47)



Problem description

- We wish to predict group membership for a number of subjects from a set of predictor variables.
- The criterion variable (also called grouping variable) is the object of classification. This is ALWAYS a categorical variable!!!

For a number of subjects from a set of predictor variables that means greater variables are very independent variables so with the help of these independent variables I want to predict group membership okay whether let say this for x will fall in to the category of loan paying category or we will not fall into the category of loan paying category so whether is a loan payer or not payer for example, the criteria variable is thus object of classification right so this is always a categorical variable.

As I have mentioned if you can see so it is always a categorical variable 0 no 1 yes right or could be 2 whatever, let us take this case example we want to know whether.

(Refer Slide Time: 07:35)

Example

- We want to know whether somebody has lung cancer. Hence, we wish to predict a yes or no outcome.
- Possible predictor variables: number of cigarettes smoked a day, coughing frequency and intensity etc.

Somebody has lung cancer or not right so there are two possible outcomes yes he has a lung cancer or know he does not have lung cancer so the possible predict variables in this case which I have taken is if you can see is like num how did I take this possible predictor variables that is simple from your maybe past experience earlier research that has been done in this area of study or maybe your own understanding so the possible predicted variables for example in this case are number of cigarettes, smoke per day.

(Refer Slide Time: 09:09)

Example

- We want to know whether somebody has lung cancer. Hence, we wish to predict a yes or no outcome.
- Possible predictor variables: number of cigarettes smoked a day, coughing frequency and intensity etc.

So they how many cigarettes the person is smoking per day is one of the important predicted variables. Second is what is the coughing frequency how in many times is if coughing right, so that would indicate whether he is got a lung cancer or not, right. The intensity of the coughing so is the coughing very highly intensive you can categorize maybe you can give a score to that, may be in a score of maybe 1 to 7 or 1 to 5 like in a likelihood scale or something that is saying yes.

Well, this is what he is doing or if you can measure the maybe the sound that is making the decibel of the sound that also could be an indicator I do not you can take any way right. So to measure this whether he is having a lung cancer or not we can use this variables the predicted variables to decide whether the person has got lung cancer or not.

But the question is how is this going to help us, now before this let me tell you with a simple common sense we can all understand that a person can only you can only predict the future when you have a past in a knowledge that means I should be having certain data for example of people who had lung cancer and how was their smoking habit.

How was their coughing frequency and how was their intensity if I knew this earlier then I could possibly create you know an equation and find out okay, whether if a new patient is coming by looking at this variables the scores or the coefficient of this variables like coughing frequency, intensity and cigarettes I can maybe predict okay well this man is maybe possible a patient of lung cancer or he is not, right, okay.

(Refer Slide Time: 10:07)

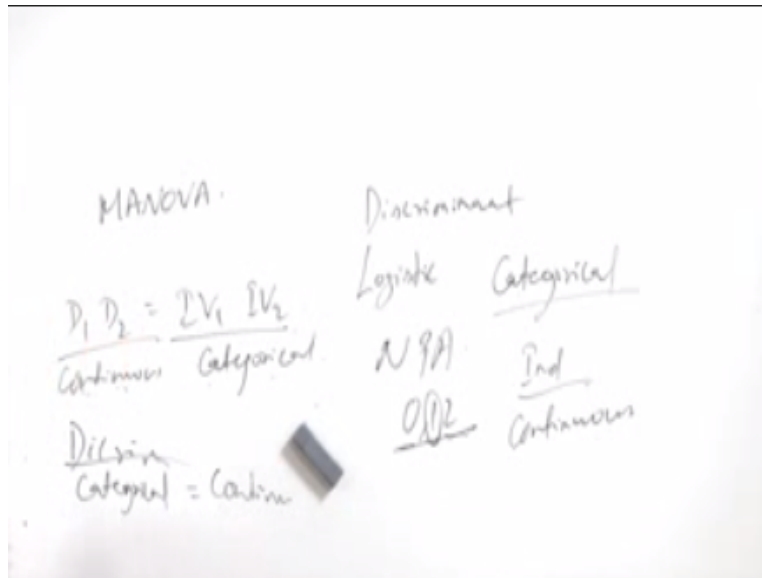
Basics

- Used to predict group membership from a set of continuous predictors
- Think of it as MANOVA in reverse – in MANOVA we asked if groups are significantly different on a set of linearly combined DVs. If this is true, than those same “DVs” can be used to predict group membership.

APPLIED ONLINE EDUCATION COURSE

So sound the basics if you look at you know as it says used to predict group membership from a set of continuous predictors right that is so we have discussed. Now the beauty is that if you remember another technique which we are discussed was called MANOVA right, an extension of ANOVA, now MANOVA if you remember what where we are doing in MANOVA we were basically trying to understand how to or how groups are behaving right, when you have multiple dependent variables and multiple independent variables. Let us say or if not multiple independent variables at this 1 but multiple dependent variables.

(Refer Slide Time: 10:51)



So when you had D1, D2 and you had some kind of IV1, IV2 right, so the relationship that we were making but only condition was that these were at that time the categorical variables right, and these were the continuous variables so if you understand, if you look at it MANOVA is nothing but the reverse of discriminate analysis or discriminate analysis is nothing but the reverse of MANOVA right.

So in MANOVA you have the continuous dependent variables and your categorical independent variables and in discriminate you have categorical independent variables dependent variables and continuous independent variables right, so that is the difference, okay. So as it says in MANOVA we ask if groups are significantly different on a set of linearly combined DVs, so you had linearly combined dependent variables.

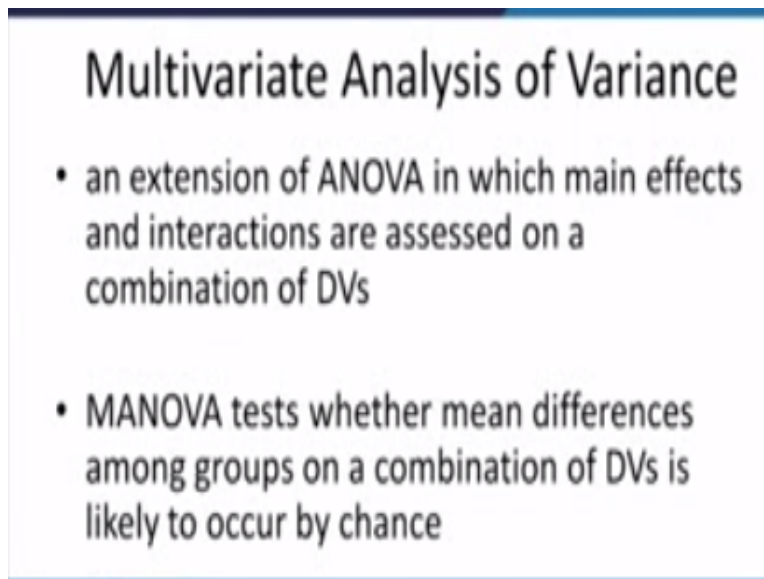
Now this dependent variables could have a high co-linearity could not have it is not advisable to have a high co-linearity between them. So anyway this could it was helping us to predict group membership right. So as I have written MANOVA tests whether means difference among groups on combination DVs likely to occur by chance or would it occur every time that means it is not a chance right.

So and you will also measuring the impact of interactions in ANOVA and MANOVA the most important thing was you are effecting the impact of interaction, now inter actions I had already explain that means let us say a b are effecting c so a is affecting c b is effecting c but when let us say let us understand this way so a b are two variables effecting c right so suppose a is effecting c

that thing is not significant be effecting c is also not significant but however there is a possibility that a cross b is effecting c and this is significant.

So that is what helps MANOVA and ANOVA helps us to understand okay so I do not want to speak too much on MANOVA now because already we have discus this things right I am straight away getting in to the discriminant analysis overview.

(Refer Slide Time: 13:22)



Multivariate Analysis of Variance

- an extension of ANOVA in which main effects and interactions are assessed on a combination of DVs
- MANOVA tests whether mean differences among groups on a combination of DVs is likely to occur by chance

So what is it? Why we use it?

(Refer Slide Time: 13:25)

Discriminant Analysis Defined

Multiple discriminant analysis . . . is an appropriate technique when the dependent variable is categorical (nominal or nonmetric) and the independent variables are metric. The single dependent variable can have two, three or more categories.

Examples:

- Gender – Male vs. Female
- Heavy Users vs. Light Users
- Purchasers vs. Non-purchasers
- Good Credit Risk vs. Poor Credit Risk
- Member vs. Non-Member
- Attorney, Physician or Professor

© 2008 SPSS INCORPORATED

So as you can see it is an appropriate technique when the dependent variable is categorical and then dependent variable are metric the single dependent variable can have two three or more categories as I said in logistic regression we generally talk about two categories only 0 and 1 right yes or no but in case of discriminant analysis you have even four categories also but beyond a certain number of categories when you create it becomes complicated it becomes a very complex model right.

So for examples I have given at the side if you see gender is he a male or female? Heavy users versus light users purchasers versus non purchasers good credit risk versus poor credit risk right member is a member or a non member attorney, physician or professor three categories right.

So the point is when you are in life coming to such a situation where you have to discriminate or you have to classify them in to different categories there it becomes very important and this tool is a very powerful tool right. So how does the dependent variable z is equal to credit risk.

(Refer Slide Time: 14:38)

Discriminant Analysis

Dependent Variable (Z) = Credit Risk

(Favorable vs. Unfavorable)

Independent Variables:

X₁ = income

X₂ = education

X₃ = family size

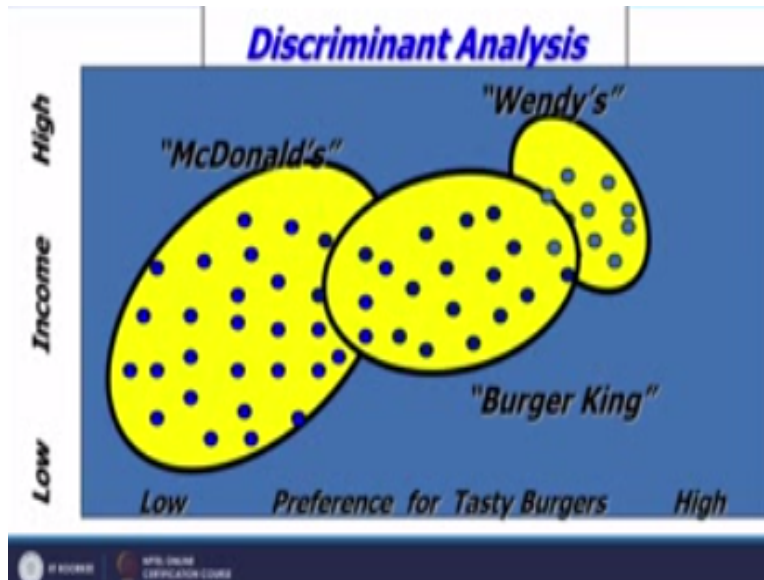
X₄ = occupation (dummy variable)

X₅ = ??

Now for in this case and independent variable is for example this is a slight case we have taken income, education, family size, occupation where occupation has been measured as a dummy right, so dummy is also very important case which is more or less use in regression which dummy variable regression so dummy variable coating is method where that means just you take the presence or absence of a item right if it is present then it is one and wherever it is not present it is 0 as good as that right.

So by taking this can be find out any relationship now once we have found this relationship then after that can I use this relationship as a predictive model in the later on stages okay. now if you see for example this is the case.

(Refer Slide Time: 15:28)



Now there are three groups right the McDonald's the Burger King and the Wendy's now preference for tasty burgers is low and high and income of people is low and high so is there any way is that help this is that explaining us can it be possible that we can this dots are nothing but the respondents, so the respondents can they be actually aggregated or clearly differentiated in to different groups on basis of their income and their preference for burgers.

Now this is just to show right now one thing that is very important you need to remember is like discriminate also analysis also had a very large similarity with you know analysis of variance right. As I said inverse what is that basically it measures basically within group variances the within group within group variances divided by the total variance right so when you do this, this value is nothing but if you see in a software or technique so this variances right it measures the variances so when two groups when the overlap each other so what it means is when the within group variances w by let say $30/T$ is very less or is sufficiently less.

Then we would say that when the wills λ is les in size you know that means there is a clear cut distinction between the groups right the higher willslambda tells you that there is a lot of overlap that is lot of overlap may be something like this right but if it is not then it would suppose the wills λ is poor then the overlap is only very less right.

(Refer Slide Time: 17:51)

<i>Survey Results for the Evaluation* of a New Consumer Product</i>				
<i>Purchase Intention</i>	<i>Subject Number</i>	<i>X₁ Durability</i>	<i>X₂ Performance</i>	<i>X₃ Style</i>
<i>Group 1</i>				
<i>Would purchase</i>	1	8	9	6
	2	6	7	5
	3	10	6	3
	4	9	4	4
	5	4	8	2
<i>Group Mean</i>		7.4	6.8	4.0
<i>Group 2</i>				
<i>Would not purchase</i>	6	5	4	7
	7	3	7	2
	8	4	5	5
	9	2	4	3
	10	2	2	2
<i>Group Mean</i>		3.2	4.4	3.8
<i>Difference between group means</i>		4.2	2.4	0.2

So that is what it actually explains in this you can see there is a sizable distinction between the two three groups Mc Donald's Burger king and Wendy's right so the question here is another question that is we have taken is on basis of purchase intention and certain variables like durability performance and style right so the numbers are just kept this side please you to understand this is subject number durability this is the durability.

And performance style I am sorry performance x2 this is x2 and this is x3 all measured in a scale of 0 to 10 okay all measurements scalar is 0 to 10 now what it does basically now here we calculate nothing but the discriminate score we calculate something call the discriminate score now discriminate score is Z is equal to A+ like an regression equation $w_1 x_1 + w_2 x_2$ it goes on right.

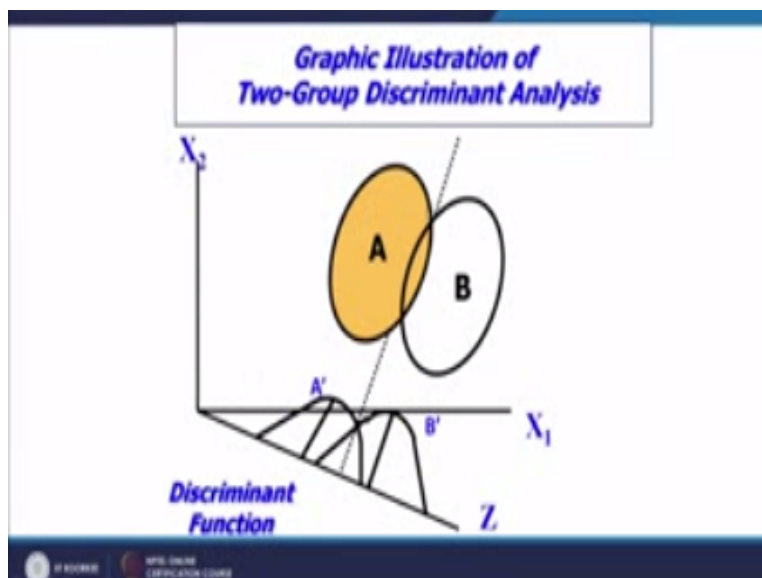
So once we have this discriminate score with us it helps us to identify whether the new person would fall into which category right so this is the survey as the for the evaluation of a new consumer product right the other groups are would purchase right and would not purchase so the group 1 are people who would purchase the item and the group 2 are the people who would not purchase.

Now if you look at the durability or the values of what they have given the people who would purchase how importance they offered to durability what is the importance they have offered to performance what is the durability importance they have offered to style right and similarly the people who are not going to purchase what is their score right so the difference between the two

group means right If this are the group means so this was the group means 7.4 and this 3.2 so the difference is given here 4.32.

Similarly 2.4 and 0.2 right so this helps us right it helps us to create kind of prediction tools say this is the characteristics of the people right, then where do you, for taking incorporating these values right. We can say if the person who is coming next, and we asked him what is your score that you will give on durability performance in style. So by taking his opinion in score we can predict whether this man would purchase or not purchase, just imagine once if you can find out this man would purchase or not purchase. How easy it is becoming for the marketer, so this is what I was talking about.

(Refer Slide Time: 20:39)



So if you look at A and B the two groups would purchase and would not purchase, there is very small overlap. That means in technical term I would say the wills γ is a very low value right. So the low value means, within group variances/ total group variances is quite low. I have explained it earlier also, within group is nothing but you can understand, in my way I have explained it has a variances. No let see the two sets.

(Refer Slide Time: 21:21)



So the objectives of the discriminate analysis, the research design assumptions, what assumptions we should have estimation interpretation and validation right.

(Refer Slide Time: 21:27)

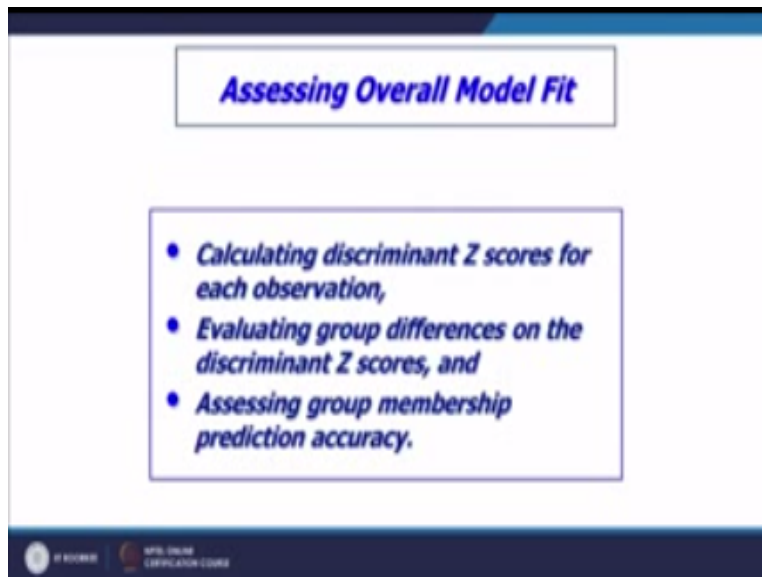


So I am not taking all but I am taking through some of the really important once, first of all how you should approach, the most common approach you have done is, suppose you have a large number of let say data which is continuous in nature. So if you have a matrix scale responded as it saying, you should convert into non metric categories. Now how is that for example, let say people with 3 income groups let say upto 0 to 5 lakhs.

5 lakhs to 10 lakhs and 10 lakhs above there are three categories are there, so just to make it one, 0m to 5 is 1, 5 to 10 is 2, and similarly 3. So by doing that it becomes easier for you to convert that in a categorical scale right. And other we use is called the polar extreme approach right. polar extreme approach are something where only two groups are used and the middle one is excluded, that means it is only the 0 or 1.

That means you do not have the 5 to 10 lakhs group in the earlier case, so it is 0 ,2 let say 10 and 10 and above okay.

(Refer Slide Time: 22:45)



Assessing Overall Model Fit

- *Calculating discriminant Z scores for each observation,*
- *Evaluating group differences on the discriminant Z scores, and*
- *Assessing group membership prediction accuracy.*

© 2019 WUOLAH EDUCATION CORP

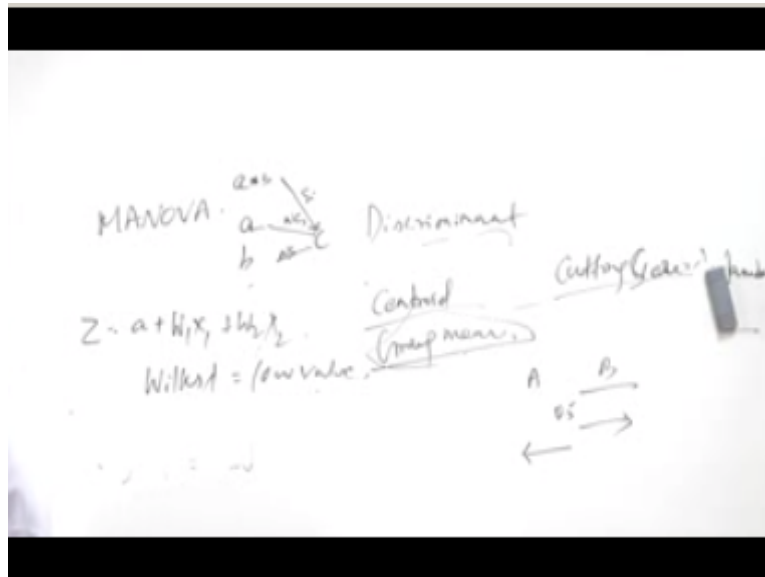
So you have the calculating discriminative Z scores right which I said and then you have to something else is called a cut off score, we develop a cut off score. Now once you have the three things you have to remember, so once you have the Z scores, then after this you create, you calculate the centroids or the group means, this is important. What are centroids or the group means? When the large number of values you have, you try to calculate the group mean right of each group.

Now by having this is also called as centroids, if any book you see do not get confused, centroids or the group means it helps you to identify how far is the particular score from the group mean, so if it is very far or it is very close, that helps you to decide whether it is, it will be falling into the particular group or not okay.

So it helps in as I said discriminative analyses are very practical significance it helps you to predict and then find out how accurately you can predict basically okay so one more thing when I am talking about group mean centroid or you know and z score is the another thing which is important for you to understand that is there is something called a cutting score now cutting score or cut of score so what is this cutting score or cut of score the cutting score or cut of score is nothing.

But you have to find a value to see truth what was happening you have created two groups right now the question is if a new person is coming you need to find out whether the person will fall into which group whether group a or group b now to fall into group a or group b there has to be a kind of a score if you so that if it is below so let us say the particular cutting score then it is in let us say group b let us say I am just telling you suppose particular value it is let us say 0.5.

(Refer Slide Time: 25:12)



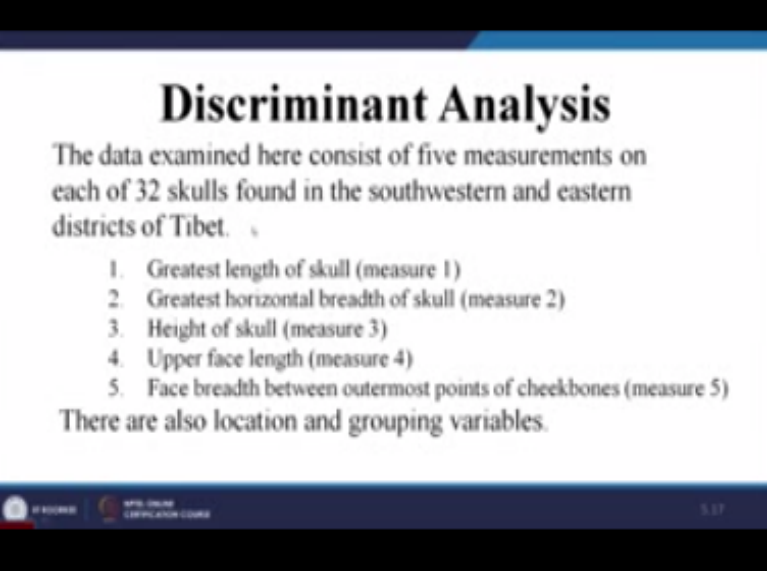
So if it is less than 0.5 let us say this group if it is more than 0.5 suppose let us say this group okay so how do you find this cutting score is a very important parameter so this cutting score is nothing but cutting you can say $z = \frac{n_b \bar{x}_b + n_a \bar{x}_a}{n_a + n_b}$ now what do I mean by this when I am saying this z cutting score is equal to number of samples or respondent in the group a multiplied by the centroid or the group mean right of group b similarly number of members in the group b into the centroid of group mean of the group a divided by the total sample size.

Now this is generally this is utilized in case of unequal groups when two groups do not have an equal number of respondents or equal size right however but if you have equal let us say group size then it will become simply $\frac{z_a + z_b}{2}$ right okay

So this is the case which I have brought to explain I hope by now you have understood that descriptive analysis there are few things to understand that it helps you to predict whether in new member will fall into which group will it fall into the passing group or the failing group will it

get into you know get into so whatever right so it helps you to basically predict so what is the next outcome whether by looking at certain continuous independent variables okay so this is the case which I have brought as an example from.

(Refer Slide Time: 27:17)



Discriminant Analysis

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

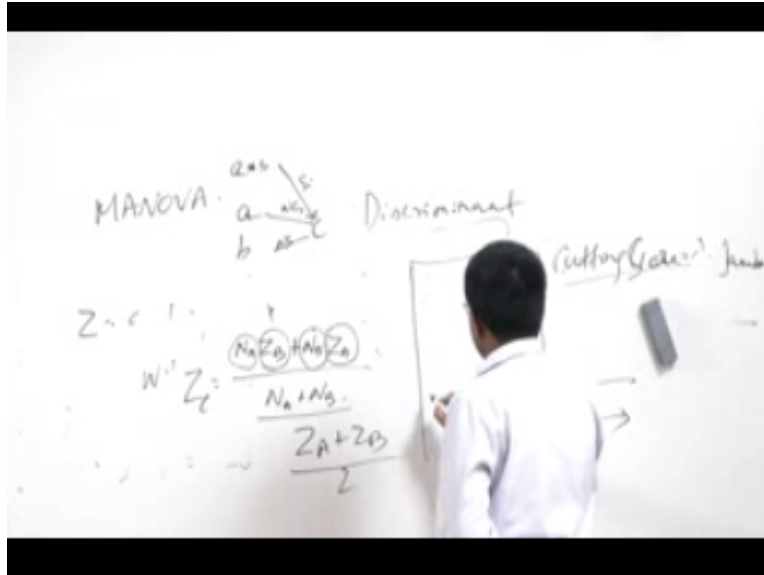
1. Greatest length of skull (measure 1)
2. Greatest horizontal breadth of skull (measure 2)
3. Height of skull (measure 3)
4. Upper face length (measure 4)
5. Face breadth between outermost points of cheekbones (measure 5)

There are also location and grouping variables.

© 2012 BY THE UNIVERSITY OF CHICAGO
3/17

The net I have brought so the data examined consist of five measurements on each of 32 skulls before that before the let me although you will do this may be in the coming session I will also explain this same I will continue with this so what happens is as I said descriptive analysis is helps you to predict right but there is also something like you know validate the discriminative analysis now how do you validate the descriptive analysis to validate the descriptive analysis what you do the simplest mechanism is very simple is that whatever your sample is right your data set whatever you have you divide the data set into two parts right one you use it for the study the other you use it has a hold out sample we say hold out you know group now what is this hold out group now after when you that means let us say suppose this is my let us say this is my entire data set

(Refer Slide Time: 28:19)



So I am just dividing the data set into two parts part a and part b this is my hold out sample right whatever I will do I will test I will do the test on this data right and then I will again run the same analysis on this data and I will compare the results if the results are coming more or else the same then I would say my results are validated right so this is the very simple mechanism to do it there is nothing science in that too much of science in that you just have to break the data into two parts that's all right.

So what I will do is I will explain you how to conduct descriptive analysis through with the help of us also right and how to interpret the results of descriptive analysis I hope at the moment the theory I very clear and you have understood while descriptive analysis or for logistics also as I said both can be used for the same purpose only difference is being that descriptive analysis is little more sensitive to the data assumptions of normality.

And all logistic is more robust but the point is logistics generally works with only two categories of variables where is descriptive value is validity more than two categories of variables up to 4 even 5 also okay 5 is too much maximum 4 so what I will do is in the coming session we will get into descriptive analysis through the help of an example okay thank you so much for this session.

For Further Details Contact

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

E-Mail Ecellitrke@gmail.com etcell@itr.ernet.in
Website: www.itr.ac.in/centers/ETC. 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 Coordinators

Prof B.K Gandhi

An Educational Technology Cell

IIT Roorkee Production