

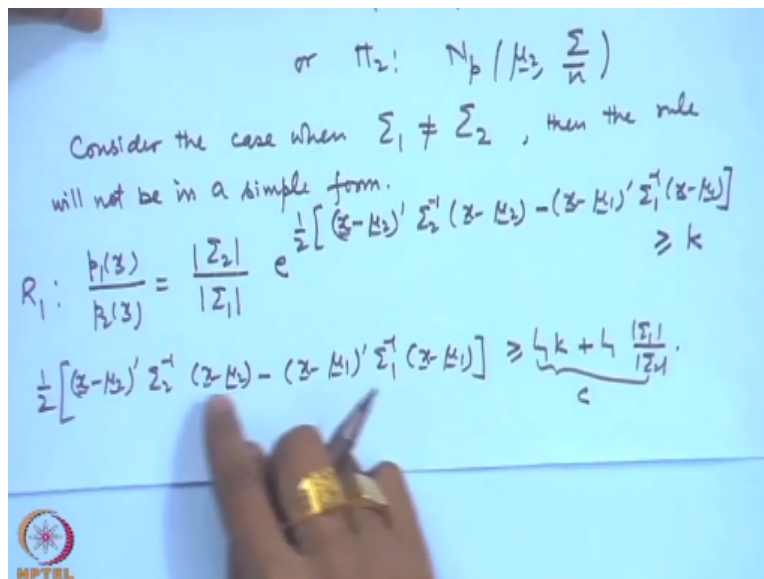
**Statistical Methods for Scientists and Engineers**  
**Prof. Somesh Kumar**  
**Department of Mathematics**  
**Indian Institute of Technology – Kharagpur**

**Lecture – 25**  
**Multivariate Analysis – X**

In the previous lecture, I have discussed the problem of classification of an observation into 2 multivariate normal populations, and we had made the assumption that all the parameters are unknown. I discussed one case in detail in which I assumed the multivariate normal population  $\mu_1$   $\sigma_1$  and  $\mu_2$   $\sigma_2$ , that means covariance matrix was assumed to be equal and known. In that case, we were able to derive the distribution of the discriminant function and the probabilities of misclassification.

That means the exact form of the rule was quite convenient to obtain. I also discussed the case when  $\sigma_1 \neq \sigma_2$  and in that case we do not have a linear discriminant function if we use the same methodology.

**(Refer Slide Time: 01:17)**



The distribution will be dependent upon the central and north-central chi-square distributions and it will be somewhat complicated. The form of the rule can be obtained but suppose we want to study the properties or we want to derive actually the minimax rule here, then that is going to be much more complicated compared to the case of  $\sigma_1 = \sigma_2$ . However, in most of the

practical situations it may actually happen that the parameters of the populations are not known.

I discussed the case of say disease etc., so from the experience of the medical practitioners, they may actually identify that this disease has that this parameter vectors and covariance matrix and similarly the other. But there can be various other problems where it is simply a problem of classification. For example, land area, economic conditions and various kinds of things. In that case, low-income group, high-income group and so on.

So, in those cases parameters although we may specify that it is a multivariate normal distribution but we may not be able to say what are the parameters of the distribution. In that case, we consider the problem by substituting the estimates of the parameters in the discriminant function. This procedure was initially proposed by Fisher in 1936 and he called it a linear discriminant function. Basically, he used the same one which I have described in the previous one but he substituted the estimates.

**(Refer Slide Time: 03:02)**

Lecture 25

Suppose the problem is to classify an observation  $X$  into one to two multivariate normal populations when the parameters of the population are not known

$$\pi_1: N_p(\mu_1, \Sigma), \quad \pi_2: N_p(\mu_2, \Sigma)$$

In this we must have some information on the populations in form of samples (called training samples)

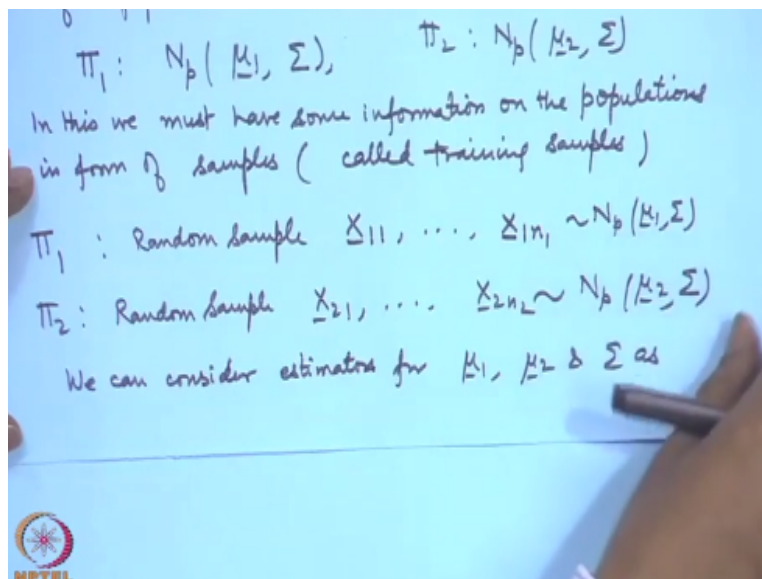
$$\pi_1: \text{Random sample } X_{11}, \dots, X_{1n_1} \sim N_p(\mu_1, \Sigma)$$
$$\pi_2: \text{Random sample } X_{21}, \dots, X_{2n_2} \sim N_p(\mu_2, \Sigma)$$

So, let me specify the problem and then, suppose the problem is to classify an observation  $X$  into one of 2 multivariate normal populations when the parameters of the population are not known. So, we are having  $\pi_1$ , say that is  $N_p(\mu_1, \Sigma)$  and  $\pi_2$  is  $N_p(\mu_2, \Sigma)$ . So, in this case, we must have some information on the populations in form of samples. These are actually called training samples.

So, for example from  $\pi_1$  we consider a random sample, say  $X_{11}$  and so on,  $X_1$  we may consider equal sample sizes are in unequal sample sizes, so these cases can also occur. So, we can consider unequal sample sizes. So, this is from NP  $\mu_1$   $\Sigma$ . Once again you note here that although parameters are unknown but I have assumed the covariance matrix to be common. There can be another case when that is uncommon.

And again you can see that there will be complications as in the case of known parameter problem and from  $\pi_2$  we have, so this is from  $X_{21}, X_{22}$  and so on  $X_2$  and 2. This is the sample from second population. So, when we have this data, we can easily consider the maximum likelihood estimators or we can consider unbiased estimators. We can look at the sufficient statistics. So, this problem is well studied in the estimation theory. So, I will not dwell too much into this and I will simply write the estimators.

**(Refer Slide Time: 06:38)**



So, we can consider estimators for  $\mu_1$ ,  $\mu_2$  and  $\Sigma$  as  $\mu_1$  head= $\bar{X}_1$ , that is actually  $1/n_1$ ,  $\Sigma = \frac{1}{n_1} \sum_{j=1}^{n_1} (X_{1j} - \bar{X}_1)(X_{1j} - \bar{X}_1)^T$  and  $\mu_2$  head= $\bar{X}_2 = 1/n_2 \sum_{j=1}^{n_2} X_{2j}$ ,  $j = 1$  to  $n_2$ . In fact, for  $\Sigma$  we can consider separately and then since  $\Sigma$  is common and we write the joint likelihood function etc.

**(Refer Slide Time: 07:01)**

© CBT  
I.I.T. KGP 2

$$\hat{\mu}_1 = \bar{x}_1 = \frac{1}{n_1} \sum_{j=1}^{n_1} x_{1j}, \quad \hat{\mu}_2 = \bar{x}_2 = \frac{1}{n_2} \sum_{j=1}^{n_2} x_{2j}$$

$$\hat{\Sigma} = \frac{1}{(n_1 + n_2 - 2)} \left[ \sum_{j=1}^{n_1} (x_{1j} - \bar{x}_1)(x_{1j} - \bar{x}_1)' + \sum_{j=1}^{n_2} (x_{2j} - \bar{x}_2)(x_{2j} - \bar{x}_2)' \right] = S$$

We use these estimates in the discriminant form:  $U$ , then we get

$$W = \underbrace{X' S^{-1} (\bar{x}_1 - \bar{x}_2)}_{\downarrow} - \frac{1}{2} (\bar{x}_1 + \bar{x}_2)' S^{-1} (\bar{x}_1 - \bar{x}_2) \dots (1)$$

Fisher's Linear discriminant fn.  
(1936).

NPTEL

So, the sufficient statistics basically reduces. Basically, we can consider then pooling and if you want consider unbiased estimators, then it will become  $1/(n_1+n_2-2)$ ,  $\sum_{j=1}^{n_1} (x_{1j} - \bar{x}_1)(x_{1j} - \bar{x}_1)'$  +  $\sum_{j=1}^{n_2} (x_{2j} - \bar{x}_2)(x_{2j} - \bar{x}_2)'$ . Actually, this one will be unbiased if you consider  $1/(n_1+n_2)$ , then that will be the maximum likelihood estimator. So, for large sample it will not make any difference whether you take this.

So, now if you consider the discriminant function that I introduced in the previous lecture, this  $X' S^{-1} (\mu_1 - \mu_2) - \frac{1}{2} (\mu_1 + \mu_2)' S^{-1} (\mu_1 - \mu_2) \geq \log K$ . So, this part if we substitute the estimates of  $\mu_1$ ,  $\mu_2$ , etc. then we get the form as. So, we use these estimates in the discriminant form. In fact, you can look at the definition of  $U$  which I gave here which was basically the left-hand side of this rule, that is this particular term.

So, this you can be considered as, then we get let us call it  $W$ , so  $W = X' S^{-1} (\bar{x}_1 - \bar{x}_2) - \frac{1}{2} (\bar{x}_1 + \bar{x}_2)' S^{-1} (\bar{x}_1 - \bar{x}_2)$ , so this I call  $S$ , okay. This term I call  $S$ ,  $\bar{x}_1 - \bar{x}_2$ , you can actually match here term by term.  $X'$ , this I am substituting  $S$ , then  $\mu_1 - \mu_2$  I am writing as  $\bar{x}_1 - \bar{x}_2$ . This is  $\mu_1 + \mu_2$ , so this will become  $\bar{x}_1 + \bar{x}_2$  and  $S^{-1} (\bar{x}_1 - \bar{x}_2)$ .

This function is actually because the right-hand side does not depend on the  $X$  value. So, this will be same for all observations which we want to classify. So, this one is the Fisher's linear discriminant function which he proposed sometime around 1936.


(Refer Slide Time: 11:19)

We use these estimates in the discriminant form. we get

$$W = \underline{x}' S^{-1} (\bar{x}_1 - \bar{x}_2) - \frac{1}{2} (\bar{x}_1 + \bar{x}_2)' S^{-1} (\bar{x}_1 - \bar{x}_2) \dots (1)$$

↓  
Fisher's Linear discriminant fn.  
(1936).

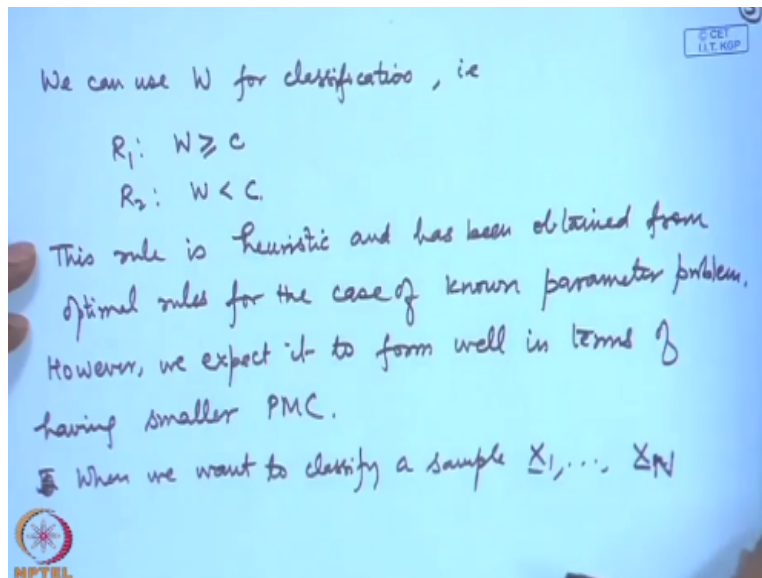
This fn. has greatest variance between samples.



This function has greatest variance between samples. So, this can be considered as the classification criteria because earlier we have used  $U$  as the classification criteria because we are getting the terms like  $U > K$ ,  $U < K$ . There it was actually proved that it is one of the base rules and therefore it falls in the class of admissible rules and therefore it is desirable and we could actually choose a minimax choice there.

Now, unlike that this one has not been derived in that fashion, because the main reason is that  $P_1$  and  $P_2$  are not completely known here. We have actually substituted the estimates but we can expect that this will actually be having in the same way as the previous one.

(Refer Slide Time: 12:30)

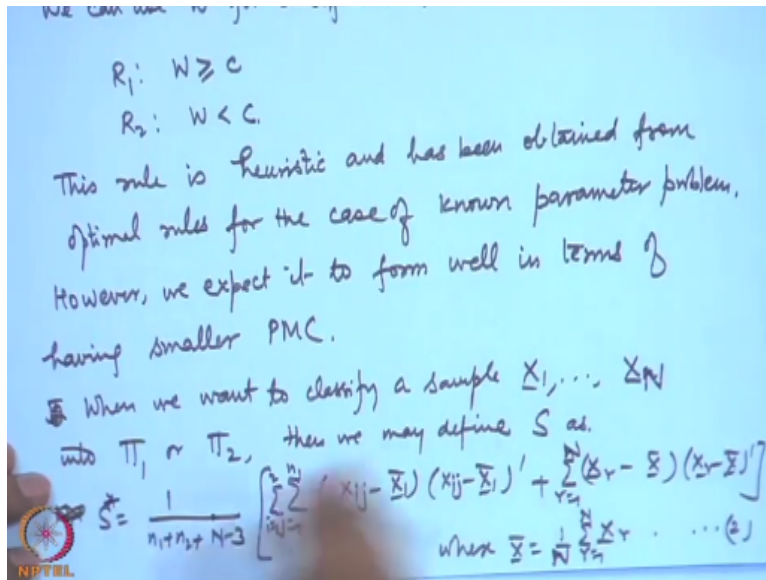


So, let me can then write here. We can use  $W$  for classification, i.e., the region  $R_1$  if  $W$  is  $\geq$  some  $C$  and  $R_2$  is if  $W$  is  $< C$ . So, as I mentioned here that we actually do not have the multi-property in the same way as we proved in the known parameter case but since we are directly substituting good estimates of the parameters, therefore we expect that this rule will also be good rule in terms of having smaller probabilities of misclassification.

Now, another problem which will be immediately coming into mind, that means in place of one observation, we will have to classify several observations, that means a sample like in the previous case I mentioned. Now, if it is one of the 2 things. For example, here  $X$  is there. Now, in place  $X$  you have say  $X_1, X_2, X_n$ . Now, in that case also the coherent matrix will be sigma. Now, one thing can be there if it is from first one then the mean is  $\mu_1$  and otherwise it is  $\mu_2$ , so that part is not known.

But for sigma we can actually make use of this sample also, the third sample because when you write the joint density this will be added there. So, in place of pooling of 2 here, we can actually add the third one also. So, that gives you a higher level you can say accuracy for the estimation of sigma. So, we consider this problem also when we want to classify a sample say  $X_1, X_2, X_N$ . Let me change this  $n$  to capital  $N$  here just to discriminate because there I am using a small  $n_1, n_2$ .

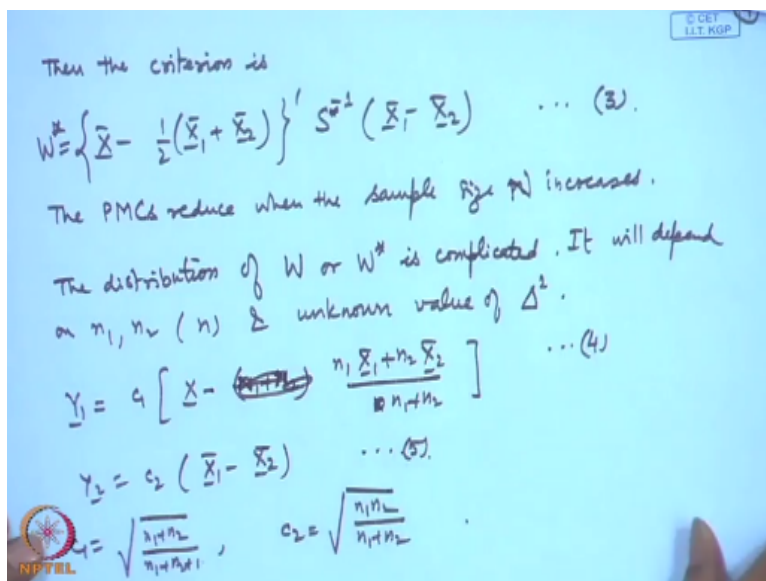
**(Refer Slide Time: 15:41)**



Suppose, we want to classify a sample this into  $\pi_1$  or  $\pi_2$ , then we may define  $S = 1 / (n_1 + n_2 + n - 3)$ , that is equal to  $\sum_{j=1}^{n_1} (X_{1j} - \bar{X}_1)(X_{1j} - \bar{X}_1)'$  and basically second term will also come, let me put it here  $I = 12 + \sum_{r=1}^n X_r - X_r - X$  transpose,  $r=1$  to  $n$ . Here this  $X$  bar is actually  $1/N \sum_{r=1}^n X_r$ ,  $r=1$  to  $n$ . Then, the criteria that will come here, so actually you can mark these 2 terms here.

You can take come as inverse  $X_1 - X_2$  because that is in both the terms. So, you are getting  $X - 1/2 X_1 + X_2$  prime.

**(Refer Slide Time: 17:36)**



So, let me give it a new name, say let me call it  $S^*$  a star here. Then  $X - 1/2 X_1 + X_2$  bar

transpose  $S^{-1} (\bar{X}_1 - \bar{X}_2)$ . General comment is that the probability of misclassification reduces when the sample size increases because the behaviour of  $\bar{X}$  will be approaching more towards the true mean, that means if true mean is  $\mu_1$ . It will approach towards  $\mu_1$  and if true mean is second one, then the true mean will be  $\mu_2$  and it will approach towards that.

So, therefore what will happen that the discrimination will be much superior because of the strong law of large number, the sample mean converges to the population mean and therefore it will be really coming out nicely there. The distribution of the criterion, that is this term, so basically this is  $W$  here, the distribution, let me call it  $W^*$ . So, we have actually either  $W$  or  $W^*$  here, that is this term which I can write as again  $(\bar{X}_1 - \bar{X}_2)^T S^{-1} (\bar{X}_1 - \bar{X}_2)$ , okay.

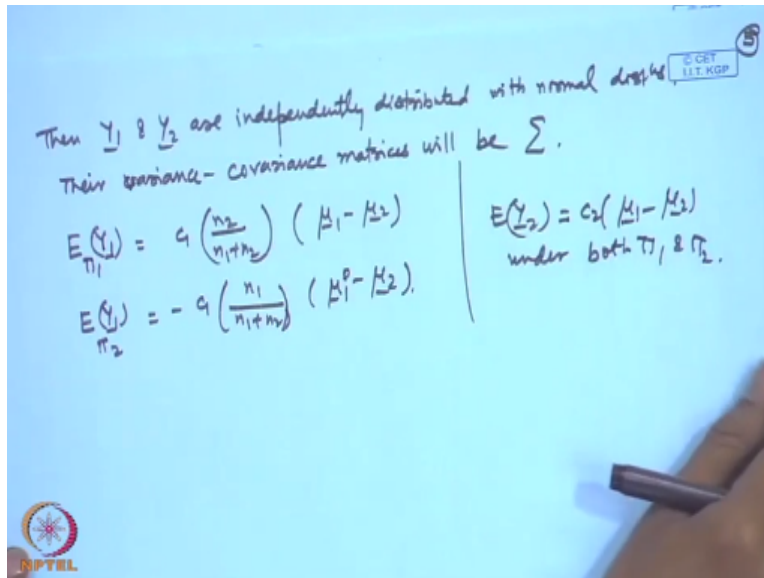
So, this  $W$  or  $W^*$ , the distribution of that will be really complicated. The distribution of  $W$  or  $W^*$  is complicated. Now, in the case of known parameters, we saw that whatever population is there, that means whether it is  $\mu_1$  or  $\mu_2$ , we got it as the univariate normal distribution and the means and the variances were easy. Actually, we got it as  $\frac{1}{2} \Delta^2$  and  $\Delta^2$  and in the second case we got as  $-\frac{1}{2} \Delta^2$  and  $\Delta^2$  was known.

Now, in this particular case it will not be simply that. It will depend upon because that is unknown here. So, it will depend upon the unknown parameter. It will depend on  $n_1$ ,  $n_2$  and of course  $n$  also and unknown value of  $\Delta^2$ , that is  $(\Delta^2)$   $D^2$  term here. I will give some representations here and in fact some work has been done by various authors on the distribution of  $W$  and  $W^*$  etc. So, let me just mention briefly some of these facts here.

One representation is given in this particular fashion that we can consider say  $Y_1$  vector as  $C_1 (\bar{X}_1 - \bar{X}_2)$ . Basically, I can write it like this  $\frac{n_1 \bar{X}_1 - n_2 \bar{X}_2}{n_1 + n_2}$ , let me call it  $Y_1$  and  $Y_2 = C_2 (\bar{X}_1 - \bar{X}_2)$  and here I am choosing  $C_1 = \sqrt{\frac{n_1 + n_2}{n_1 + n_2 + 1}}$  and  $C_2 = \sqrt{\frac{n_1 n_2}{n_1 + n_2}}$ .

**(Refer Slide Time: 22:26)**

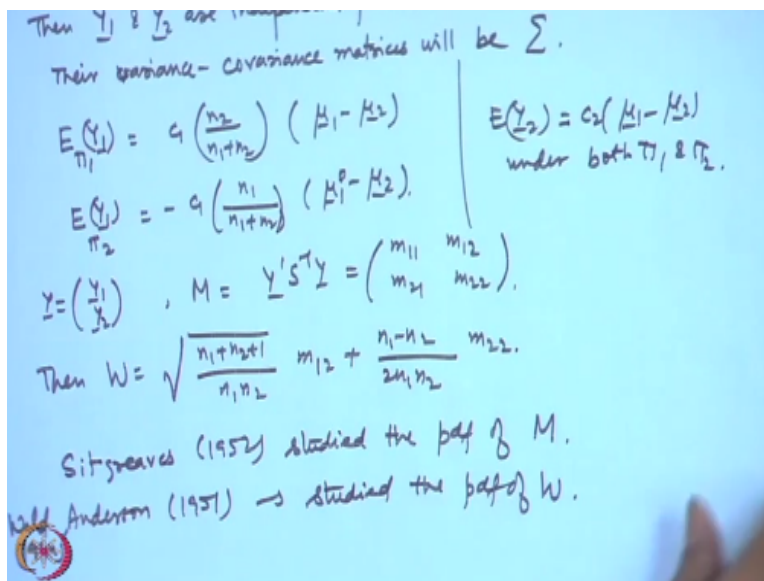




Then, we can say, actually the way this is defined here we can say that  $Y_1$  and  $Y_2$  are independently distributed with normal distributions and both will have the covariance matrix equal to  $\Sigma$ , their covariance matrices will be  $\Sigma$ . Then expectation of  $Y_1$  under  $\pi_1$ , that will be equal to  $C_1 \frac{n_2}{n_1+n_2} \mu_1 - \mu_2$  and expectation of  $Y_1$  under  $\pi_2$  that will be equal to  $-C_1 \frac{n_1}{n_1+n_2} \mu_1 - \mu_2$ . The term  $Y_2$  does not involve  $X$ .

So, this expectation will be same for both. That means if I consider expectation of  $Y_2$  that is always equal to  $C_2 \mu_1 - \mu_2$  under both  $\pi_1$  and  $\pi_2$ .

**(Refer Slide Time: 24:24)**

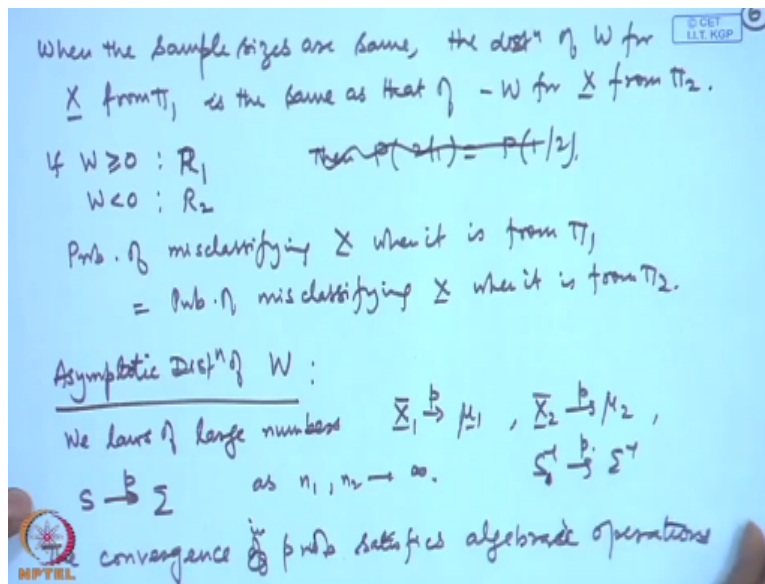


Let us define say  $Y = Y_1/Y_2$  and  $M = Y' S^{-1} Y$ , that is  $m_{11}, m_{12}, m_{21}, m_{22}$ . Then  $W$

can be written as square root of  $n_1 + n_2 + 1/n_1 n_2 m_1^2 + n_1 - n_2/2n_1 n_2 m_2^2$ . So, the density of  $M$ , it has been studied by Sitgreaves in 1952, the density of  $M$ ; then Anderson in 1951, Wald of course in 1944, they have studied the density of  $W$ .

One special case that is when the sample sizes are equal, then some simplification does occur, because actually in that case this term will simply vanish. If this term vanishes and also you can look at here, this will become half, half, etc. So, things will become much simpler. Here also it will become symmetric. So, if you look at this one then also symmetry will occur here. So, we will discuss this case separately.

**(Refer Slide Time: 26:00)**



When the sample sizes are same, the distribution of  $W$  for from  $\pi_1$  is the same as that of  $-W$  for  $X$  from  $\pi_2$ . So, if we consider  $W \geq 0$  as the region of classification into  $\pi_1$  and  $W < 0$  as the region of classification in this one, then  $P_2$  given 1 =  $P_1$  given 2. Basically we can say that the probability of misclassifying  $X$  when it is from  $\pi_1$  is equal to probability of misclassifying  $X$  when it is from  $\pi_2$ .

So, this case is a simplified version actually and basically what is happening if you use this, then basically you are considering a good rule there because the probability of misclassification for both the population will be same. That means whether it from  $\pi_1$  or from  $\pi_2$ . So, this rule will be somewhat alright. Exact distribution is quite complicated but if we look at the expressions

here since the expression of the criteria is involving.

For example, you look at  $W$  here, so if we consider the strong law of large number then  $\bar{X}_1$  will converge to  $\mu_1$ ,  $\bar{X}_2$  will converge to  $\mu_2$ ,  $S$  will converge to  $\Sigma$  inverse and so on. We can also look at the for example weak law of large numbers, that means convergence in probability. See, these convergences in probability or convergence is almost surely, that is strong law and weak law, they will actually satisfy algebraic operations.

For example, we may consider the summation, the products, multiplications, they will be invariant, that means this will converge to exactly  $U$ . If I take the previous  $U$  which I gave when the known parameters were there, then this will actually converge to  $U$  in probability. In fact, it will converge strongly, that means it will converge with probability one. Now, if that happens that means for large sample sizes, the distribution of  $W$  is almost the same as the distribution of  $U$  and in that case the probabilities of misclassification etc. have been already considered.

So, that is fine here. So, let us see here asymptotic distribution of  $W$ . By laws of large numbers,  $\bar{X}_1$  converges to say  $\mu_1$  in probability  $\bar{X}_2$  converges to  $\mu_2$  in probability and  $S$  will converge to  $\Sigma$  in probability as  $n_1, n_2$ , to infinity and of course  $S$  inverse will converge to  $\Sigma$  inverse in probability. The convergence in probability satisfies algebraic operations.

**(Refer Slide Time: 31:05)**

$$\text{Hence } S^{-1}(\bar{X}_1 - \bar{X}_2) \xrightarrow{p} \Sigma^{-1}(\mu_1 - \mu_2)$$
$$\text{2 } (\bar{X}_1 + \bar{X}_2)' S^{-1}(\bar{X}_1 - \bar{X}_2) \xrightarrow{p} (\mu_1 + \mu_2)' \Sigma^{-1}(\mu_1 - \mu_2)$$

as  $n_1, n_2 \rightarrow \infty$ .

So the limit distribution of  $W$  is the same as that of  $U$ .

Hence you will have  $S^{-1}(\bar{X}_1 - \bar{X}_2)$  converging to  $\sigma^{-1}(\mu_1 - \mu_2)$  and  $S^{-1}(\bar{X}_1 + \bar{X}_2)$  converging to the corresponding term  $\sigma^{-1}(\mu_1 + \mu_2)$  as  $n_1, n_2 \rightarrow \infty$ . So, the limiting distribution of  $W$  is the same as that of  $U$ . So, basically we can say that for large sample size, we are behaving as if the parameters of the populations are known.

And therefore under the probabilities of misclassification will be similar to the probabilities of misclassification as is the own parameter case. So, that means we are not going to do much worse here. It will be basically almost the same here. There are other derivations of the criteria based on regression criteria. Then, there is also a criteria called likelihood ratio criteria which is the same as basically the likelihood ratio test procedure.

In fact, if you remember we have written  $P_1(X)/P_2(X) > K$ . If you remember Neyman Pearson Lemma, for simple hypothesis testing problem that means if you have  $H_0: P_1$  and  $H_1: P_2$ , then basically your hypothesis testing problem is also decided on the basis of that. What is the most powerful tests? The most powerful test is based on that is acceptance and rejection regions are based on the ratio  $P_1$  and  $P_2$ .

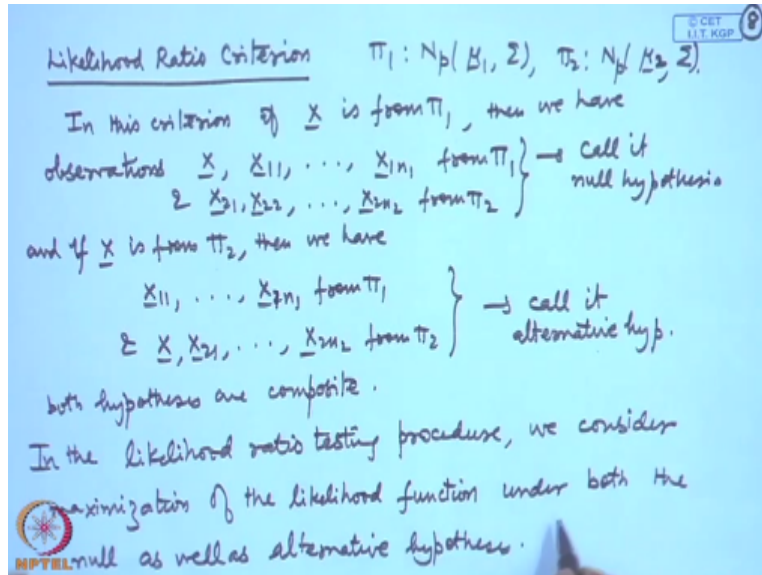
There we write  $P_0$  and  $P_1$ , so it is the same thing basically. That is, we have  $P_1/P_0 > K$  for  $P_1/P_0$  greater than something. So, it is the same thing. Now, in the likelihood ratio procedure of course in that one we just tried the thing and we consider the size of the test and based on that we consider the maximization of the power. So, the constant  $K$  was chosen subject to that condition. The reason is that the problem of testing of hypotheses is interpreted differently.

Because there we have the probabilities of type I error and type II error and we cannot actually consider those probabilities equal kind of thing. But in this particular case it is a different matter. There we had fixed like probability of type I error say  $\alpha$  and we try to minimize the probability type II error, here that criteria is not done. Rather we are looking at the probabilities of misclassification.

And in the base rule, we are actually considering the minimisation of the probability of

misclassification and in the minimax rule we are considering the equating of those things. So, it does not mean that we are proceeding in the same way, although the form is the same. Another procedure in the testing was the likelihood ratio, that procedure can also be adopted. So, I will derive the classification procedure based on the likelihood ratio criteria.

**(Refer Slide Time: 35:10)**



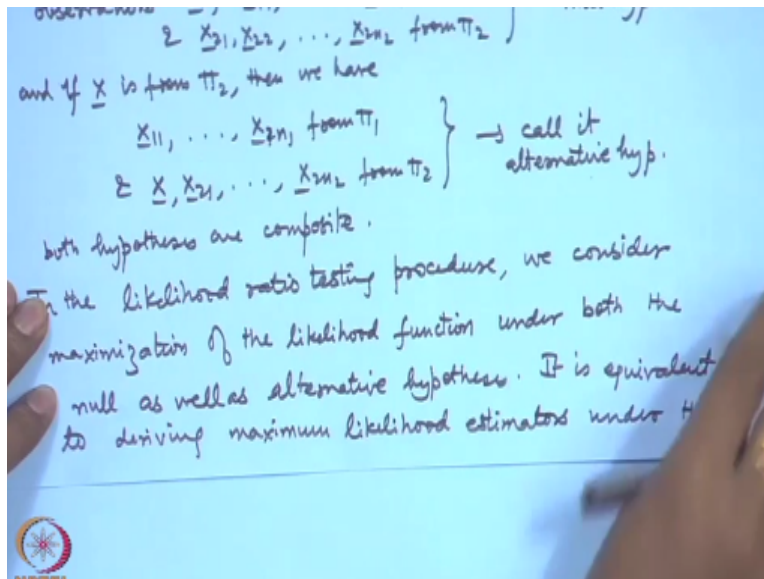
Let us look at this thing. Let me leave this here. Likelihood ratio criterion, so in the likelihood ratio criterion if  $X$  is from  $\pi_1$ , then we have observations  $X, X_{11}, \dots, X_{1n_1}$  from  $\pi_1$  and also  $X_{21}, X_{22}, \dots, X_{2n_2}$ , these are from  $\pi_2$ , so this is my null hypothesis. If  $X$  is from  $\pi_2$ , then we have  $X_{11}$  and so on...  $X_{1n_1}$  from  $\pi_1$  and  $X, X_{21}$  and so on...  $X_{2n_2}$  from  $\pi_2$ . This we call as alternative hypothesis. So, you can note that both hypothesis is composite. See this  $\pi_1$  and  $\pi_2$  we have already written here,  $\pi_1$  is actually  $N_p(\mu_1, \sigma)$  and  $\pi_2$  is  $N_p(\mu_2, \sigma)$ .

Here all  $\mu_1, \mu_2$  and  $\sigma$  they are unknown. In the likelihood ratio criteria, when we consider the null and alternative hypothesis and composite hypothesis, what we consider is the maximisation of the likelihood function over the null hypothesis space and the alternative hypothesis space, and then we take the ratio of the 2 likelihood functions and we consider greater than or less than.

So, let us consider this makes maximum likelihood estimation here. Basically, maximization of the likelihood function; in the likelihood ratio testing procedure, we consider maximisation of the

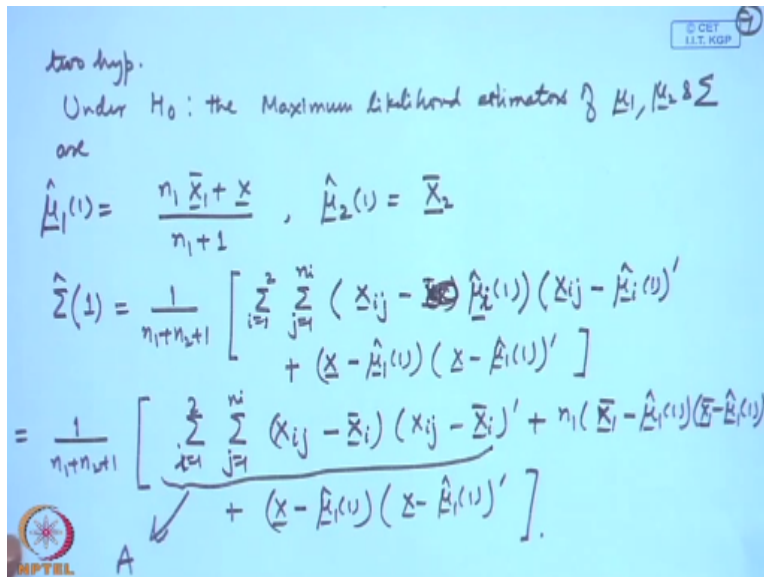
likelihood function under both the null as well as alternative hypothesis.

(Refer Slide Time: 39:30)



It is equivalent to finding maximum likelihood estimators under the 2 hypotheses. So, let us consider firstly. So, this is my  $H_0$  and this is my  $H_1$ , okay.

(Refer Slide Time: 40:07)



So, under  $H_0$ . So, you can look at this problem carefully. We are considering  $X, X_{11}, \dots, X_{1n_1}$ , this is a sample from  $N(\mu_1, \sigma^2)$ . So, when we write the likelihood function, it will be the joint likelihood function of  $n_1 + 1$  observations from this and similarly for the second one, it will be these  $n_2$  observations from this. So, we will get a combined term here and when we consider the maximum likelihood estimators.

The maximum likelihood estimators of the parameters  $\mu_1$ ,  $\mu_2$  and  $\sigma^2$ , so I will call it  $\mu_1$  head 1 =  $n_1 \bar{x}_1$ . This observation will come because you are considering the mean of these  $n_1 + 1$  observations. For the second one, it will be simply the mean of  $n_2$  observations. So, in the first case it is in  $n_1 + 1$  observation. In the second case, it is  $n_2$  observations here. So, that is this one and the sigma head, let us call it under 1 =  $1/n_1 + n_2 + 1$ .

Now, you will see that it will remain common here,  $\sigma^2 = \frac{1}{n_1 + n_2 + 1} \sum_{i=1}^2 \sum_{j=1}^{n_i} (X_{ij} - \hat{\mu}_i)(X_{ij} - \hat{\mu}_i)'$ , rather than this it will be actually the mean here, that is  $\mu_i$  head 1, right \*  $X_{ij} - \mu_i$  head 1 transpose and extra term will come here because of  $X$  here. So, if you consider  $X$ , then  $X - \mu_1$  head 1 \*  $X - \mu_1$  head 1 transpose. If we want, we can substitute these values here. Basically you can consider in the first case  $\bar{X}_1$  and in the second case  $\bar{X}_2$ , then there will be some sort of simplification here and the terms then can be written like this.

Let me express it here, this can be further written as  $1/n_1 + n_2 + 1$ . Let me write it separately, it is becoming  $\sum_{i=1}^2 \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)(X_{ij} - \bar{X}_i)'$ ,  $j=1$  to  $n_i$ ,  $i=1$  to 2. Then, there will be some extra term coming in here, i.e.,  $+n_1(\bar{X}_1 - \hat{\mu}_1)(\bar{X}_1 - \hat{\mu}_1)'$ . See, these terms that I am getting here, I can define the term called A. So, this particular term which I have written here. This term let us call it A.

**(Refer Slide Time: 46:21)**

$$\begin{aligned} \hat{\Sigma}(1) &= \frac{1}{n_1 + n_2 + 1} \left[ \sum_{i=1}^2 \sum_{j=1}^{n_i} (X_{ij} - \hat{\mu}_i)(X_{ij} - \hat{\mu}_i)' + (\bar{X} - \hat{\mu})(\bar{X} - \hat{\mu})' \right] \\ &= \frac{1}{n_1 + n_2 + 1} \left[ \underbrace{\sum_{i=1}^2 \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)(X_{ij} - \bar{X}_i)'}_{A} + n_1(\bar{X}_1 - \hat{\mu}_1)(\bar{X}_1 - \hat{\mu}_1)' \right] \\ A &= \frac{1}{n_1 + n_2 + 1} \left[ A + \frac{n_1}{n_1 + 1} (\bar{X} - \bar{X}_1)(\bar{X} - \bar{X}_1)' \right] \end{aligned}$$

So, then we can actually express it in this particular fashion as  $1/n_1+n_2+1$ . See these 2 terms again there is some sort of combining that can be done. You look at the nature here nature of the terms. Here it is  $\bar{X}_1 - \mu_1$  head,  $\bar{X}_1 - \mu_1$  head prime and here also  $\bar{X} - \mu_1$  head  $1 * \bar{X} - \mu_1$  head  $1$  prime. So, we can actually consider expansion of this and then combine this term here. If we do that, so basically if you look at this term here, then there is a  $\mu_1$  head prime which is coming here also, but here it is  $n_1$ . So, it will become actually  $n_1+1$ .

So, if you take it out and consider the division there, then we can express the terms as  $n_1/n_1+1$   $\bar{X} - \bar{X}_1$  bar,  $\bar{X} - \bar{X}_1$  bar transpose. So, you can see here that I am able to write down the maximum likelihood estimator. See this part we spent some time just to express it in a nice fashion. Otherwise, these terms are also fine but this form you can see it will be helpful because ultimately we have to write down the ratios of the terms. So, then you consider here.

**(Refer Slide Time: 48:19)**

Under  $H_1$ : the MLEs of  $\mu_1, \mu_2$  &  $\Sigma$  are

$$\hat{\mu}_1(\gamma) = \bar{X}_1, \quad \hat{\mu}_2(\gamma) = \frac{n_2 \bar{X}_2 + X}{n_2 + 1}$$

$$\hat{\Sigma}(\gamma) = \frac{1}{n_1 + n_2 + 1} \left[ A + \frac{n_2}{n_2 + 1} (\bar{X} - \bar{X}_2) (\bar{X} - \bar{X}_2)' \right]$$

The  $\frac{\hat{L}(\text{null hyp})}{\hat{L}(\text{alt. hyp})} = \left\{ \frac{|\hat{\Sigma}(\gamma)|}{|\hat{\Sigma}(1)|} \right\}^{-\frac{n_1 + n_2 + 1}{2}}$

or  $\frac{1 + \frac{n_2}{n_2 + 1} (\bar{X} - \bar{X}_2)' A^{-1} (\bar{X} - \bar{X}_2)}{1 + \frac{n_1}{n_1 + 1} (\bar{X} - \bar{X}_1)' A^{-1} (\bar{X} - \bar{X}_1)}$

Under  $H_1$ , the maximum likelihood estimators of  $\mu_1$ ,  $\mu_2$  and  $\Sigma$ , then what this will become equal to,  $\mu_1$  head 2, that will be  $\bar{X}_1$   $\mu_2$  head that will be  $n_2 \bar{X}_2 + X / n_2 + 1$  and for covariance matrix, it will become  $1/n_1+n_2+1$  and if we express in the same fashion, then it will become. Now the likelihood ratio criteria what we do, we write down the ratio of the joint likelihood functions. So, if you look at the exponent term.

In the exponent term after shifting these terms, it will actually become  $E$  to the power  $-1/2$



something, that is  $-1/2$  and  $1+$  and  $2+1$   $P/2$ . So, that we will cancel out. Now in the density function, you have in the denominator determinant of sigma. So, the determinant of sigma term that is coming there that will appear and since the estimates are there sigma head 1 and sigma head 2, so they will appear here and the power will become  $n_1+n_2+1/2$ .

So, we consider the likelihood ratio that is let me call it  $L$  head under  $H$  knot under the null hypothesis divided by  $L$  head alternative hypothesis. If we consider this, then this is equal to sigma head 2/sigma head 1 to the power  $n_1+n_2+1/2$ . Now, these expressions are already available here, i.e., sigma head 1 and sigma head 2. So, if I put it here this is becoming equal to  $1+n_2/n_2+1$ ,  $X-X_2$  transpose  $A$  inverse  $X-X_2/1+n_1/n_1+1$   $X-X_1$  bar prime  $A$  inverse  $X-X_1$  bar.

In place of  $A$ , we can actually use the  $S$  term,  $S$  term we had derived earlier. Let me just show it again. This was the  $S$  term here. If we consider this  $S$  term here  $1/n_1+n_2-2$  this term, this was the  $S$  term. So, we can actually write in terms of  $S$  because he only the divisor is coming.  $A$  I wrote as this term. So,  $A/n_1+n_2-2$  is  $S$ .

**(Refer Slide Time: 52:55)**

The image shows handwritten mathematical work on a blue background. At the top, there is a fraction representing a likelihood ratio:

$$= \frac{(n_1+n_2-2) + \frac{n_2}{n_2+1} (\bar{X} - \bar{X}_2)' S^{-1} (X - \bar{X}_2)}{(n_1+n_2-2) + \frac{n_1}{n_1+1} (\bar{X} - \bar{X}_1)' S^{-1} (X - \bar{X}_1)}$$

Below this, it says "The region of classification" and defines  $R_1: \Lambda \geq c$ . A note explains that  $c$  is a function of  $n_1$  and  $n_2$ . It then states that if  $c=1$ , the rule is called the "Maximum likelihood rule".

The test statistic is given as:

$$Z = \frac{1}{2} \left[ \frac{n_2}{n_2+1} (\bar{X} - \bar{X}_2)' S^{-1} (X - \bar{X}_2) - \frac{n_1}{n_1+1} (\bar{X} - \bar{X}_1)' S^{-1} (X - \bar{X}_1) \right]$$

Finally, the classification regions are defined as  $R_1: Z \geq 0$  and  $R_2: Z < 0$ .

So, if we use this expression, then this is same as  $n_1+n_2-2+n_2/n_2+1$   $X-X_2$  bar prime  $S$  inverse  $X-X_2$  divided by  $n_1+n_2-2+n_1/n_1+1$   $X-X_1$  prime  $S$  inverse  $X-X_1$ . So, you can consider the classification region. Suppose we consider  $R_1$ , then it is this ratio which we can call it say lambda is say  $\geq$  some value  $K$  or  $C$ , okay. This can also be shown to be if you remember your

earlier criteria that was in terms of  $W$ , so you can see that they are equivalent.

If we consider  $W$  term here  $W$  was given by  $\frac{1}{S} \frac{X_1 - X_2}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$  and this term. So, this term and this term are having some equivalence here because if you look at this greater than something and then this term will get cancelled out and you adjust this term. You take this thing come here, then this will be coming exactly of the same form. So, there is a particular case which is equivalent to  $W \geq C$  say some  $C$  star if  $n_1$  and  $n_2$  are large.

If we take  $C$  to be 1, then the rule is called ML rule, that is the maximum likelihood rule, that means we are simply considering that choice where the likelihood functions maximisation gives you the higher value. So, for example if we consider say  $Z$  is equal to say  $\frac{1}{2} \frac{X_2 - \bar{X}_2}{\sqrt{\frac{1}{n_2} + \frac{1}{n_1}}} - \frac{X_1 - \bar{X}_1}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$ . Then, basically rule is  $R_1$  region is  $Z > 0$  and  $R_2$  if  $Z < 0$ .


So, basically we can think of this as a distance. This is the estimated distance of  $X$  from  $\mu_1$  and this is distance from  $\mu_2$ . So, you are saying if the distance from  $\mu_2$  is more than the distance from  $\mu_1$ , then you put it into the first one, that is  $\mu_1$ . So, basically you can consider this as a simple distance of the observation from a given population which is based on the sample from that population. So, I think the rule is a straightforward and it is extremely heuristic rule that is coming here.

**(Refer Slide Time: 57:15)**

$$W-Z = \frac{1}{2} \left[ \frac{1}{n_1+1} (X - \bar{X}_1)' S^{-1} (X - \bar{X}_1) - \frac{1}{n_2+1} (X - \bar{X}_2)' S^{-1} (X - \bar{X}_2) \right]$$

which goes to 0 in prob as  $n_1, n_2 \rightarrow \infty$ .

So asymptotic PMC are same for W or Z.



Basically if you look at W and Z here, they are not much different if you consider say W-Z, then that is simply  $\frac{1}{2} \left[ \frac{1}{n_1+1} (X - \bar{X}_1)' S^{-1} (X - \bar{X}_1) - \frac{1}{n_2+1} (X - \bar{X}_2)' S^{-1} (X - \bar{X}_2) \right]$ . This basically converges to 0 in probability as  $n_1, n_2 \rightarrow \infty$ . So, basically, asymptotic probabilities of misclassification are same for W or Z, i.e., whether we consider W or Z, they are almost the same.

So, basically what we are saying is that we can use either of these things and these statistics you can say W and Z basically they are invariant also. If you consider say shifting. For example, you look at this Z. So, if I translate all the observations, then it will not affect here. There is no affect here. So, these rules are also translation invariant and therefore that is another plus point. In the next lecturer, I will discuss the criteria for classification into several populations and also we will spend some time on the principal component analysis and canonical correlation.

So that is the remaining topic in this thing. The problem of classification has been actually studied in great detail but in this particular course we will cover only the popular one that is based on normal distributions. There are results which are available for other populations also there are certain current work going on when there are restrictions on the parameter displaces. For example, in the 2 normal populations when you are considering.

For example, if all components of  $\mu$  are the same, so like that. There can be several cases.

Suppose, you consider 2 univariate populations,  $\mu_1$   $\sigma_1^2$ ,  $\mu_2$   $\sigma_2^2$ . Now, you may consider say some additional information in the form of say  $\mu_1 \leq \mu_2$  or  $\sigma_1^2 \leq \sigma_2^2$ . In that case, what would be the classification rule.

Similarly, we can consider exponential populations. Suppose  $\mu_1 = \mu_2$  and  $\sigma_1^2 \leq \sigma_2^2$ , in that case what would be the classification rule. So, such problems are being studied currently by various researchers. So, in this particular course, we have just given or you can say the basic criteria how we can actually derive such classification procedures and we have given some optimality criteria also.

So, in the next case I will try to wind up this particular portion that is a problem of classification and give some glimpse of 2 other problems, they are called problem of principal component analysis and the problem of the canonical correlations, so that we will be covering up next to lecture.