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

Lecture - 22
Multivariate Analysis VII

In the previous class we have discussed methods for testing of parameters for 1 multivariate normal distribution and also for 2 multivariate normal distributions. So for example if we have 1 sample from normal mu sigma distribution of p dimension then we can test about mu=mu0. We have seen that if sigma is known then the test will be based on a chi square statistic whereas if sigma is unknown then the test is based on Hotelling's T square statistics.

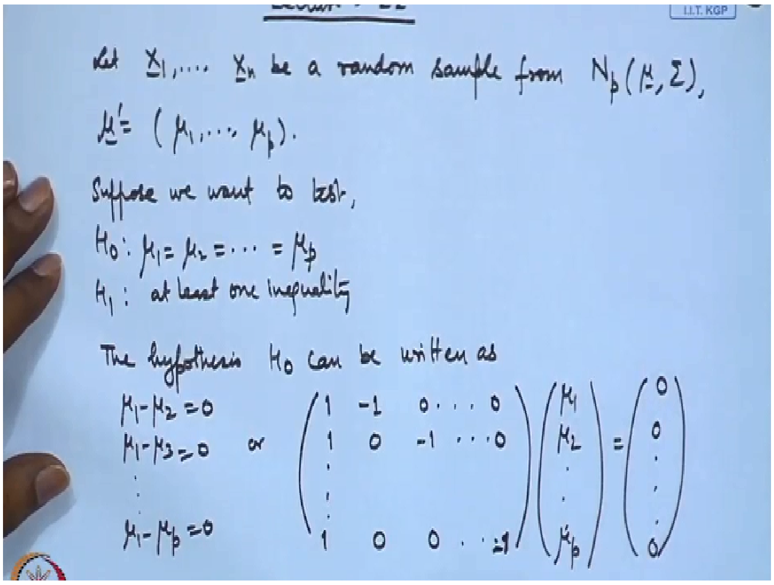
And we have shown equivalence that the value of the T square statistics can be calculated from an F distribution and the corresponding formula was given. We also discussed 2 sample problem that means we can consider testing for the equality of the mean vector of 2 multivariate normal population mu1=mu2 and again if the co-variance matrices were known the test was based on a chi square.

As well they were unknown, but equal then it was a based on a Hotelling's T square distribution. One more application of this type of testing I also showed for the linear functions of mean vector that we can consider the test for that. Now another application of this is that we may consider equality of the components themselves. Now this could be like this that.

For example, this mu1, mu2, mu p they may be denoting the characteristics of say different components of something which may have similarities. So now we would like to know whether the mu 1= m2=mu p or not that is something like a test for homogeneity. Now we know that in analysis of variance we have a test when we are considering several normal populations then it is called a one-way analysis of variance test.

But there the populations are considered to be independent that means the sampling procedure that means we are having then p independent samples. Now here by definition the samples are not independent because it is coming from a multivariate normal population.

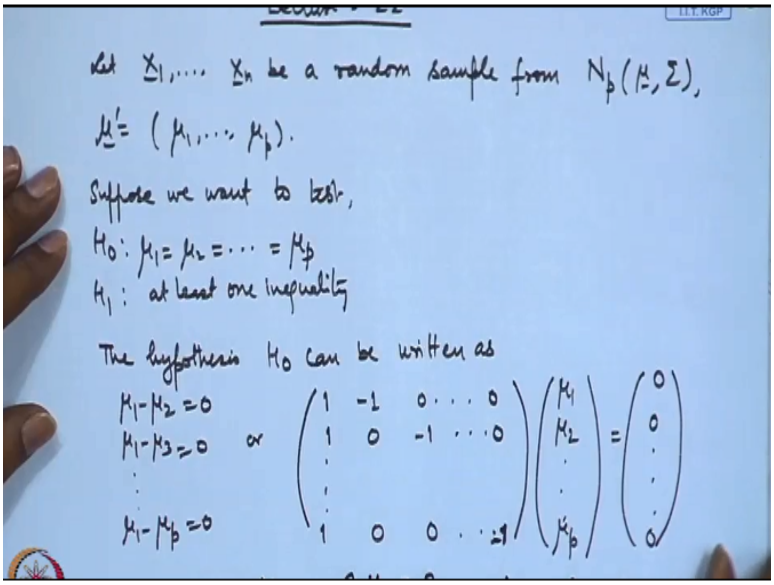
(Refer Side Time: 02:30)



So now I present a procedure for this. Let us consider a random sample x1, x2, Xn be a random sample from Np mu sigma. So as usual mu vector let me write it in the row form then mu transpose=mu1, m2, mu p. Now suppose we want to test say H0 mu1=mu2=say mu p against at least one in equality. So what we do we write it like this.

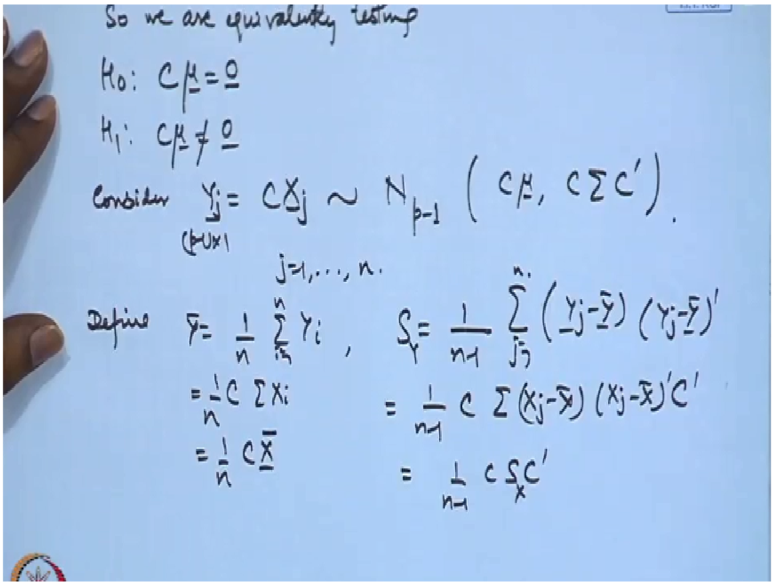
We can consider the hypothesis H0 can be written as mu1-mu2 say mu1-mu3=0 this is=0 and so on mu1-mu p=0 which we can write as say 1-1, 0, 0, 0, 1, 0-1, 0 and so on 1, 0, 0,-1* mu 1, mu 2, mu p this is=0, 0, 0.

(Refer Slide Time: 04:48)



This is we can consider it as some C matrix multiplied by mu=null where C is actually p-1/p matrix. Now you see this statement which is written in a linear form like mu 1=mu2=mu p. I can consider it as p-1 simultaneous linear functions of the mu vector of the components of mu. So we can write it as C mu=0.

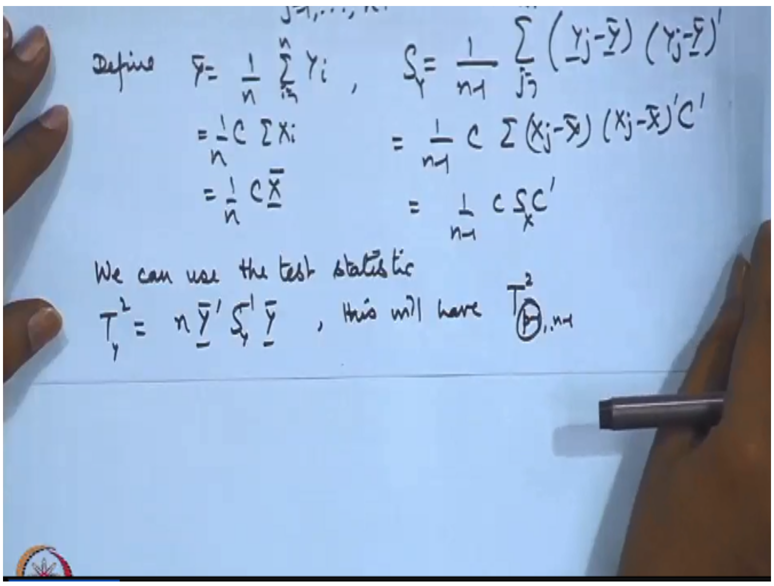
(Refer Slide Time: 05:36)



So we are equivalently testing H0, C mu=0 against H1 say C mu is not= 0. So consider the transformation say Y=CX. So this will become then p-1/1 vector then this will follow Np-1, C mu and C sigma C transpose. So let us consider yj vectors or j= 1 to n and define then y bar vector as the mean vector of yi and we can also define the variance covariance metrics based on y as 1/n-1 sigma yj-y bar yj-y bar prime j=1 to n.

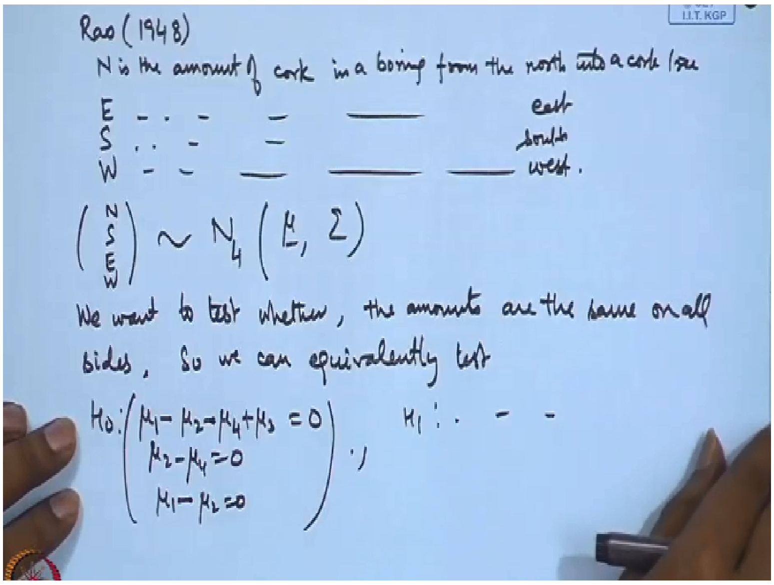
In fact, is nothing but see this one for example it is=1/n C times sigma Xi and similarly this one is 1/n-1 C times sigma xj- x bar, xj-x bar prime C prime that is actually 1/n-1 CSC prime where this S is based on x.

(Refer Slide Time: 07:28)



And this is simply so we can make use of the test statistic. Let us call it Ty square that is=n y bar prime S inverse y bar. This will have T square distribution on p-1, n-1 this is p-1 dimensional here and of course n-1 because you have n observations here.

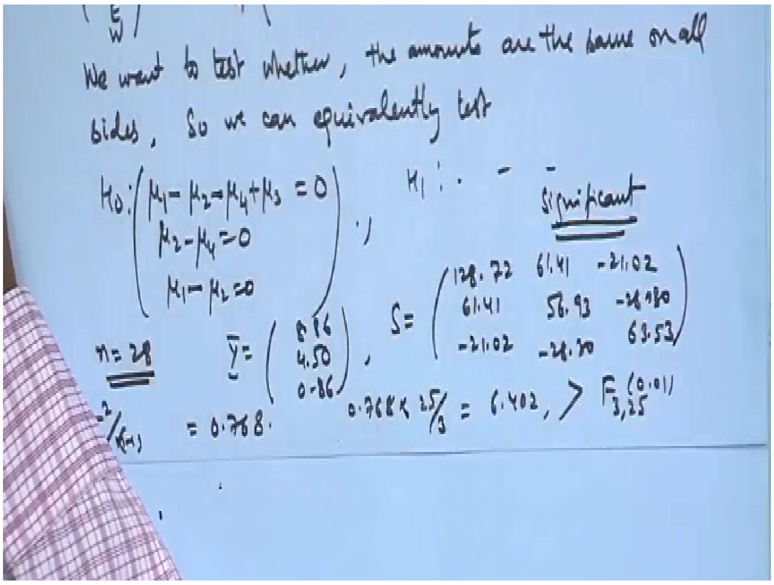
(Refer Slide Time: 08:17)



Let me give one example which is adopted from C. R Rao 1948 work and here N is the amount of I have used the terminology of that only which is amount of the cork in a boring from the north into a cork tree. And similarly you can consider E, S, W this is from East, this is from South and W is from the West and it is considered that NSEW this follows a 4 dimensional multivariate normal distribution with some mean vector mu and variance covariance matrix sigma.

And we want to test whether the amounts are the same on all sides. So we can write so as I have explained here we can use this set of hypothesis mu1-mu2=0, mu1-mu3=0, mu1-mu4=0 or we can also use say mu1-mu2-mu4+say mu3=0. Mu2-mu4=0, mu1-mu2=0 etcetera. So we can write it in any other fashion and against H1 some inequality here.

(Refer Slide Time: 10:49)

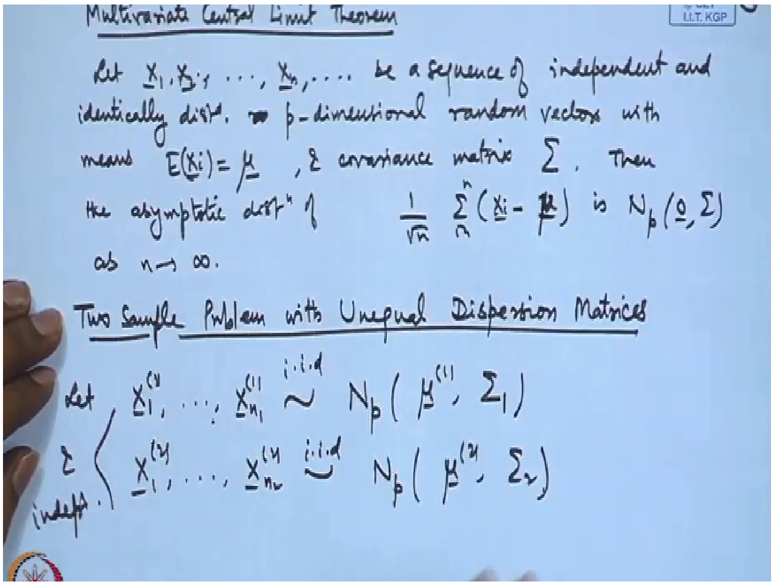


So the experiment that was conducted and reported in Rao it was having 28 observations based on that y bar was find to be 8.86, 4.50 and 0.86 and S was calculated 128.72, 61.41-21.02, 61.41-21.02, 56.93-28.30, -28.30, 63.53. So if we calculate the T square value here/n-1 that turns out to be 0.768. And if I consider compared to the F here then multiplied by 25/3 that is 6.402 and if I consider F value on 325 say at 0.01 then this is more than this.

So this turns out that this is significant. That means the amounts which are collected from all the 4 sides they vary. So this is an application of Hotelling's T square we can basically what we are showing here is that we can consider linear functions here. Now as in the case of 1 variable the importance of the normal distribution (()) (12:24) from the fact that if we consider the sums of the observation from a sample or the means of the observation from the sample then using central limit theorem we get the approximate normal distribution.

Now a similar result holds for the multivariate data also.

(Refer Slide Time: 12:48)



So we can call it Multivariate Central Limit Theorem. So that is one reason that why the methods for the multivariate normal distributions are widely applicable. So we had in the following form that let x1, x2, Xn and so on be a sequence of independent and identically distributed p dimension random vectors with means as mu and covariance metric say sigma. Then the asymptotic distribution of say 1/root n sigma Xi-mu I=1 to n.

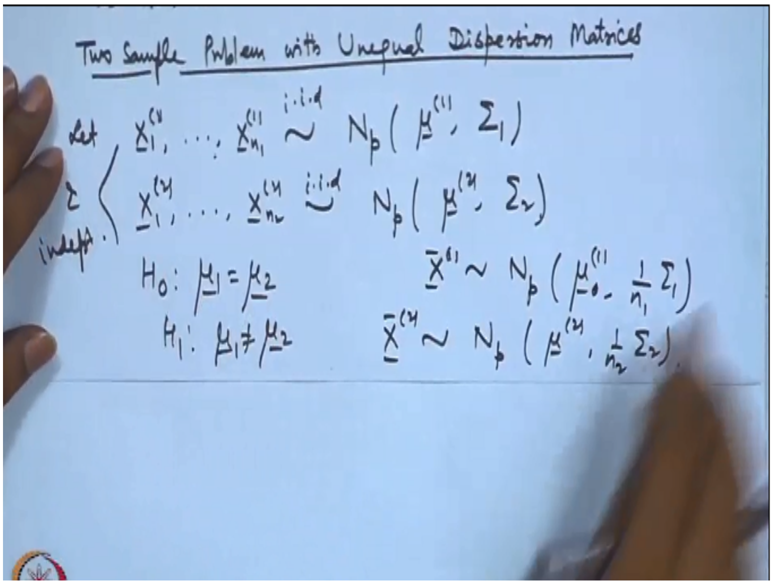
This is Np 0, sigma as n tends to infinity. So this is a version of that I have not considered divisions that means in the case of univariate we were considering deviser as sigma here, but that we are not putting here because you have a matrix here. So at the most you can consider multiplication by sigma to the power -1/2 here but that is of course easy to understand. So this type of result is helpful to establish that we can actually use the 1 sample and 2 samples procedures that we have discussed here for the multivariate normal distribution.

They will be widely applicable. Now one more case that for which in the univariate case we had some approximate procedure that was when we are considering the test for equality of variance means, but the variances are unknown as well as we do not have any other information on them like they are equal then we have a procedure which is the pooled procedure and for which I have presented the analog for the multivariate case also for the pooled Hotelling's T square.

But when they are not equal then in the case of one variable we had some approximate procedures. Now in the case of multivariate we present some procedure which is based on considering a curtailment of the observations. So let me present one procedure here. So two sample problem with unequal dispersion matrices. So we consider let x11 let us go back to the notation that I introduced earlier x11, n1 here.

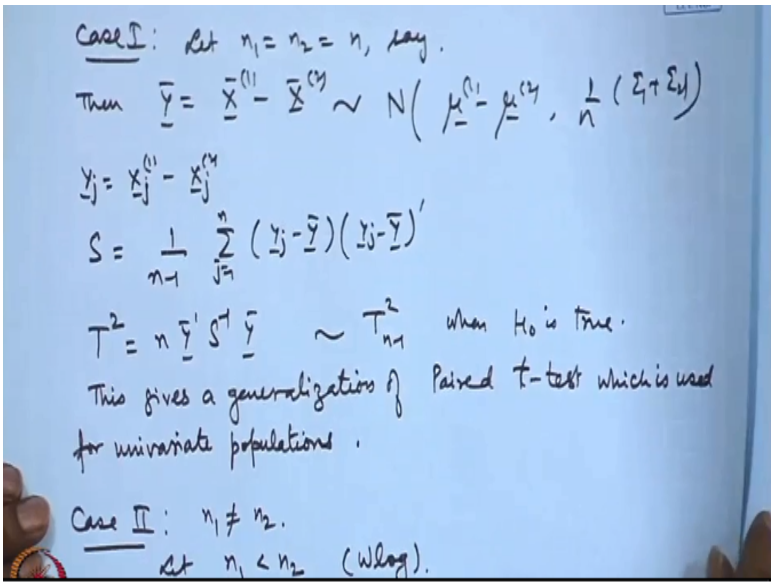
This is Np mu1 sigma 1 and another random sample say x1 2 and so on xn2 2 this is random sample from Np mu 2 sigma 2. So we have 2 independent random samples and these 2 samples are also considered to be independent here.

(Refer Slide Time: 17:00)



We are considering the test of hypothesis mu1=mu2 against mu1 is not=mu 2. Let me write the summary statistics here for example if I consider the mean. The first mean that will be Np mu1 1/n1 sigma 1 and the second mean which I call X2 that will be Np mu2 1/n2 sigma 2.

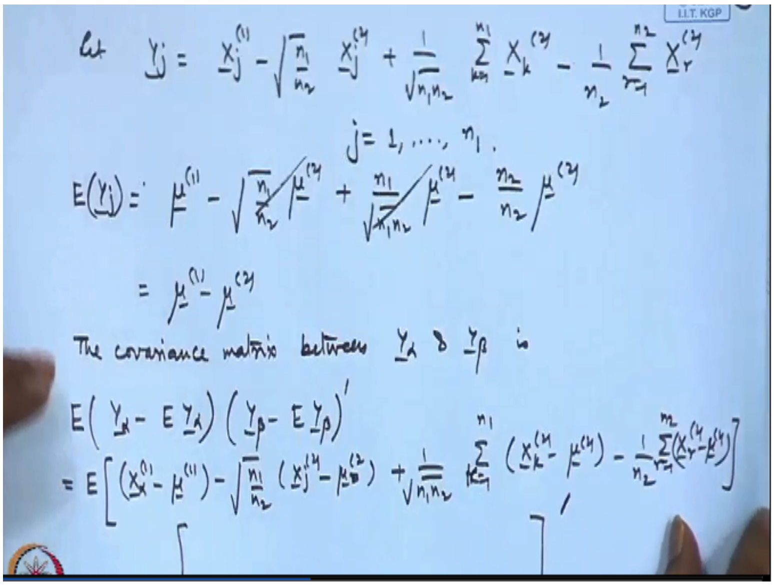
(Refer Slide Time: 17:39)



So if I consider let n1=n2=n. Then if I considered y bar that is= X bar 1-X bar 2 then that will be normal mu 1-mu 2, 1/n sigma1 + sigma2. You can see here that this because of this coefficient getting 1/n1 and 1/n2 being same I can combine this sigma1 + sigma2. And therefore I can consider here say yj=x1j-x2j. Based on this we can define S=1/n-1 sigma yj-y bar, yj-y bar transpose.

And we can consider Hotelling's T- square n y bar prime S inverse y bar. So this will follow T square on n-1 when H0 is true. So this gives a generalization of paired t test. The paired t test that we defined in the case of univariate populations which is used for univariate populations. Now let us consider the second case which is the more important one that is n1 is not=to n2. So if n1 is not= to n2 without loss of generality let us consider that n1<n2.

(Refer Slide Time: 19:58)



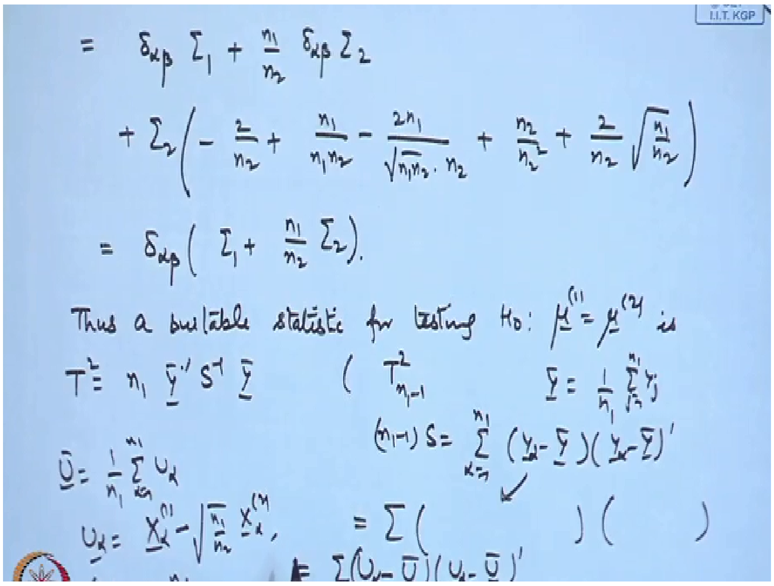
In this case let us define say yj that is= xj1-squre root n1/n2 xj 2+1/root n1, n2 sigma x k 2 k=1 to n1-1/n2 sigma r=1 to n2 x r 2. You see here that in what way we have defined see this is the observations from the first sample and here the observation from the second sample are considered here. This definition we are considering from 1 to n1. So the remaining observation that we are putting together here. Let us see the effect here.

If I consider the mean of this, then I get here the mean of the first one that is mu 1-root n1/n2 the mean of the second one that is mu2+ now here the mean of xk2 is mu2 and these are n1 observations so it becomes n1/root n1 n2 mu2-and here it will become n2/n2 mu2. You look at this terms here this term will simply get cancelled out. So we are actually getting mu 1-mu2.

That means if we base our test on the mean of yj then it will be able to test about equality of mu1 and mu2. Also let us consider the covariance matrix between say 2 observations say y alpha and y beta that is=expectation of y alpha-y beta-expectation y beta transpose. So this we expand this is= x alpha 1-mu1-root n1/n2 xj 2-mu 2+1/root n1 n2 sigma r=1 to n1 sorry this is k= 1 to n1 x k 2-mu2-1/n2 sigma x r 2-mu 2.

This is from r=1to n2 * this transpose. Now if we consider this if I consider this into the first term here then that will give me simply the first one that is sigma one the variance covariance matrix of x alpha and let us adjust the terms for the other one also.

(Refer Slide Time: 23:31)

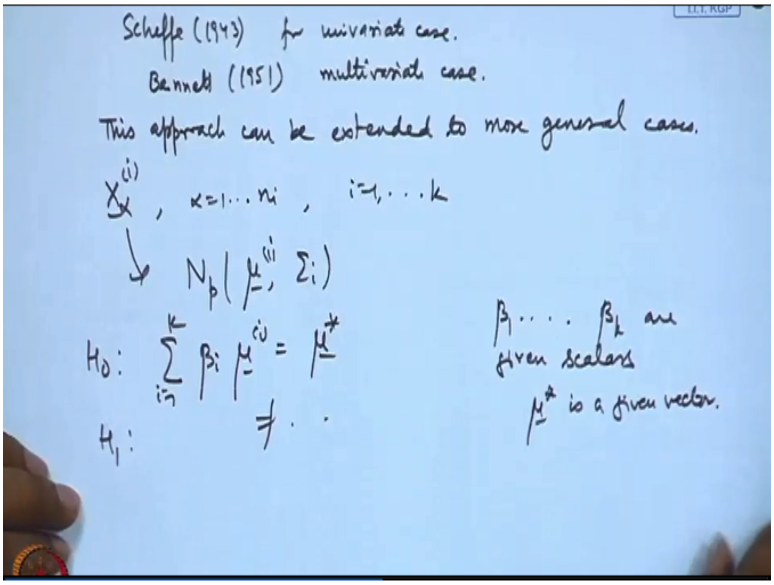


So this gives us that is= this coefficient we combine together delta, alpha, beta. See this delta, alpha, beta will be 1 when alpha=beta otherwise it is 0 sigma 1+n1/n2 delta alpha beta sigma 2+sigma 2-2/n2+n1/n1 n2-twice n1/square root n1 n2 * /n2 +n2//n2 square + 2/n2 root n1/n2. So after simplification I get simply delta alpha beta sigma 1+n1/n2 sigma 2. So based on this I can easily define suitable statistic for testing mu1=mu2.

It is based on so n1 y bar prime S inverse y bar this will have T2 on n1-1 where y bar is nothing, but 1/n1 sigma yj j=1 to n1 and n1-1 S that is nothing but y alpha-y bar y alpha-y bar transpose alpha=1 to n1. Again this can be simplified if I substitute the terms here that is if I write the full form of this y alpha and y bar here then this is actually giving us sigma U alpha-U bar U alpha-U bar transpose where U bar is=1/n1 sigma U alpha alpha=1 to n1.

And U alpha are nothing, but x alpha 1-square root n1/n2 x alpha 2. This is for alpha=1 to n1.

(Refer Slide Time: 26:24)

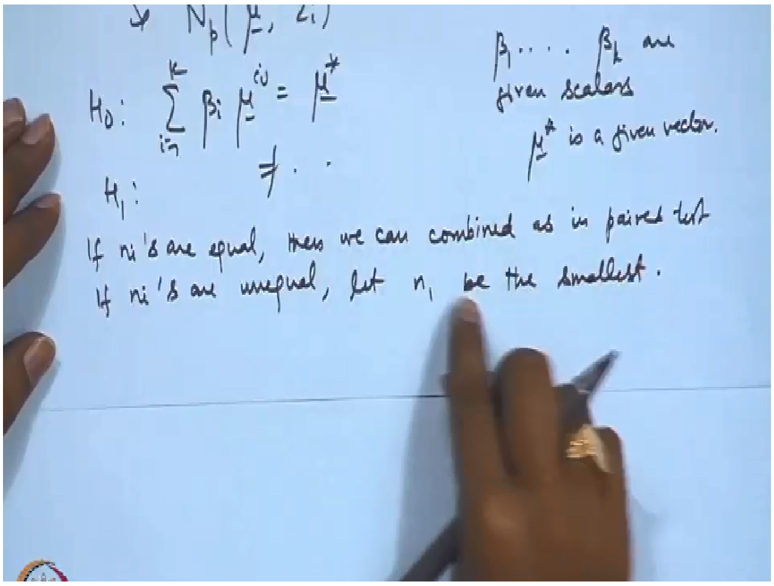


So this procedure was proposed by Scheffe in 1943 for the univariate case and Scheffe actually showed that this gives us the shortest confidence interval for the T distribution using the T distribution. Here we are actually making sacrifice of some of the observations and Benneth in 1951 he gave an extension to the multivariate case. Now one can actually consider it for several populations also when we are considering the linear combinations.

So what you will have to do you have to consider the minimum sample size of all the observations and based on that you can construct the statistics. Let me just demonstrate that thing here. This approach can be extended to more general cases. Let us consider say X alpha i or alpha= 1 to ni i= 1 to k. So these are samples are Np say mu i sigma i. So we are considering k independent samples from k Np mu i sigma i population.

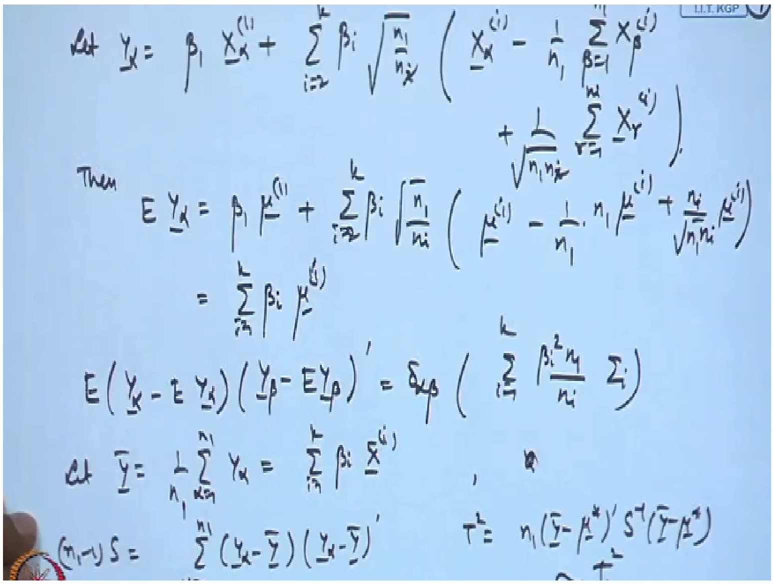
And we are considering testing for a linear combination of the mean vectors against say not equal where this beta 1, beta 2, beta k are given scalars and this mu * is given vector.

(Refer Slide Time: 28:39)



If ni's are equal, then there is no problem then we can combine as in paired test. If ni's are unequal, let n1 be the smallest and like in the previous one we consider based on n1. So again here we will do it on base of n1.

(Refer Slide Time: 29:14)

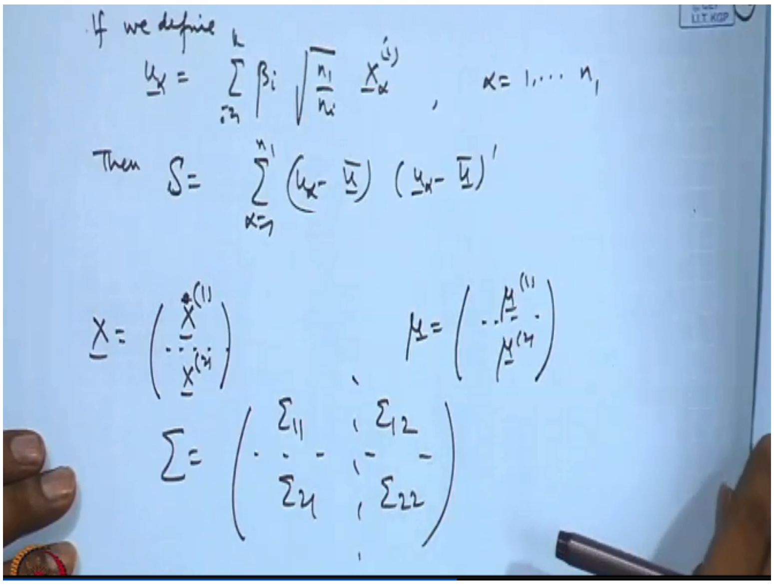


And we can define y alpha to be beta 1, x alpha 1+sigma beta I root ni/n1. This is from 2 to k and now we adjust the terms x alpha I-1/n1 sigma beta=1 to n1 x beta I + 1/ square root n 1 n 2 sigma Xri r=1 to ni. Then if we consider say expectation of y alpha then it is simply becoming beta 1 mu 1 from first term here +sigma beta i root n1/ni i=2 to k mu i-1/n1 n1 mu i+ni/ root n1 this should be ni here n1 ni and mu i.

So this term gets cancelled out n1/ni this gets cancelled here. So you get simply beta i mu i which is the desired term in the hypothesis here. So we can consider and similarly if we consider the variance covariance matrix of this based on y alpha y beta-expectation y beta transpose then that is= delta alpha beta sigma beta i square n1/ni sigma i i=1 to k. If I assume y bar as the mean of this based on n1 observations only.

And so this is-= sigma beta i x bar i where of course x bar i is the mean of the (()) (31:46) sample. And n1-S is y alpha-y bar y alpha-y bar transpose alpha=1 to n1. Then if I consider t square as n1 y bar-mu * prime S inverse y bar-mu * then that will have t square p n-1. So we can consider the Hotelling's T square test based on this here.

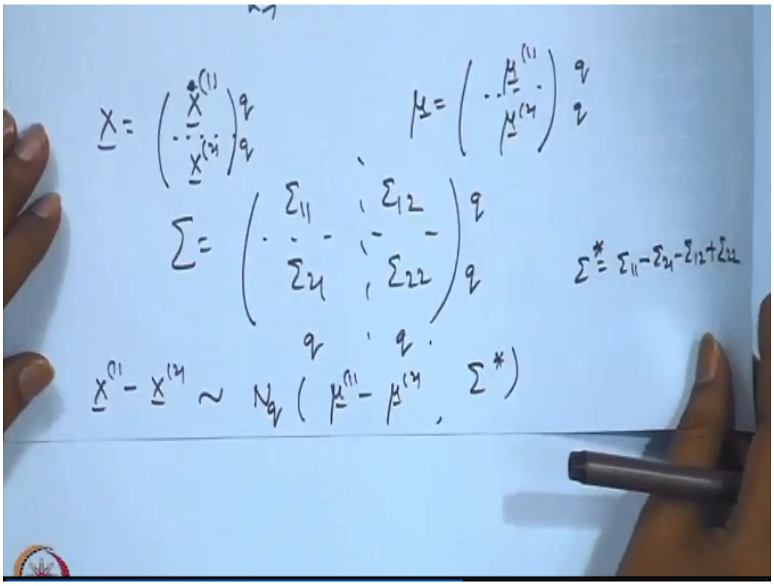
(Refer Slide Time: 32:27)



If I define say u alpha=sigma beta i root n1/ni X alpha i=1 to k for alpha= 1 to n1 then based on this S is nothing, but sigma u alpha –u bar u alpha-u bar transpose. So one can use this based on the Hotelling's T square statistics. Another problem which may also arise that I consider the 2 sub vectors of the full vector and now I want to test whether they are having equal components that means like first one I write as mu 1 mu 2 second as mu 3, mu 4 then whether mu 1= mu 3, mu 2= mu 4 etcetera.

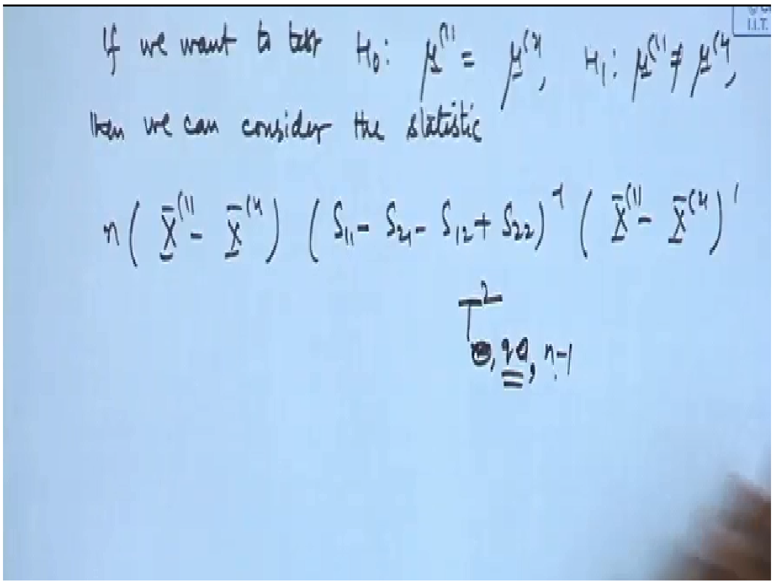
So this type of problem can also be handled using the Hotelling's T square. Let me give one example. Suppose I consider x1 and x2 here and mu=say mu 1, mu 2. So these are partitioned here, these are partitioned here and similarly the variance covariance matrix is partitioned.

(Refer Slide Time: 34:10)



So we are assuming here this as q components this as q components okay. So if I consider say x bar 1-x bar 2 then that will have q dimensional normal distribution with mean mu1-mu2 and variance covariance matrix sigma * where this sigma * can be then written as sigma 11-sigma 21-sigma 12+sigma 22.

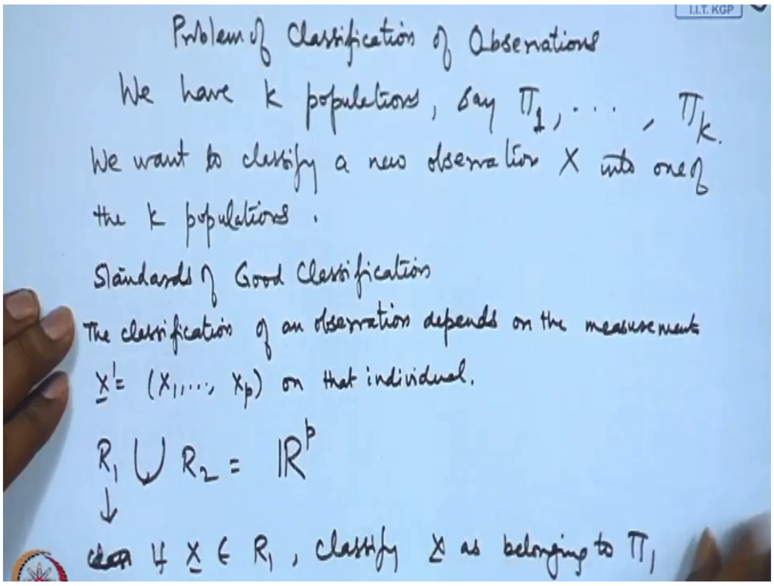
(Refer Slide Time: 34:48)



So if we want to test say H0 mu 1= mu2 against say H1, mu 1 is not=to mu 2. So we can consider the statistics n x bar 1-x bar 2 S11-S21-S12+S22 inverse X bar 1-x bar 2 transpose. So this will be based on Hotelling's T square on n-1 q-1 here sorry q and n-1 here. I have shown that various inferential problems for the mean vectors of 1 or 2 multivariate normal populations or several multivariate normal population or they can be handled using the Hotelling's T square statistics.

So there are other things which are based on the variance covariance you have something based on the Wishart distribution. However, I am not discussing that part right now because the testing for the variance covariance matrix will be somewhat little more complicated rather we move over to a more practical oriented problem which is called a Problem of Classification.

(Refer Slide Time: 36:42)



Let me introduce this problem here Problem of Classification of Observations. So quite frequently we are encountered with various kind of problem for example you consider a new entrant for example college. Now the students of the college can be described into 2 parts. One who go for an academic career and another who go for corporate job. Now based on the previous data we have the distribution of the 2 student performances.

Now when new student is considered then to which group he would belong to. Now this kind of problem can be considered in a more general setting. We have k population say pi 1, pi 2 pi k. We want to classify a new observation X into one of the k populations. So broadly speaking this is the problem of classification. Now here there can be several variations for example we may know the forms of pi 1, pi 2, pi k.

For example, this could be normal say mu 1 sigma 1, normal mu 2 sigma 2 normal mu k, sigma k and now we have another vector say x new observable we want to classify where it will belong to. Here it could be that mu 1 sigma 1 mu 2 sigma 2 mu k sigma k are known. There could be another problem when these parameters are unknown in that case we need some sort of observations from each of the populations.

Because then we will need to estimate mu 1, mu 2 mu k and sigma 1, sigma 2, sigma k. These are called training samples. There can be yet another type of problem when the forms of pi 1, pi 2 pi k are completely unknown. So in that case we have non parametric procedures. So let me introduce this problem that means what are the procedures and in what way we can study this.

So what are the standards of good classification? In a very rough way simple way we can say if we classify an observation into one of the population then either it is a correct classification or it is a incorrect classification. So a criteria for checking the goodness of the classification procedure could be the probability of incorrect classification that means we call it the probability of misclassification.

So if the probability of the misclassification remains low then it is a good procedure. So it is something like in the testing of hypothesis problem where we accept or reject the hypothesis based on the sample. Now the hypothesis could have been true and we would have rejected it and the hypothesis could have been false and we could have accepted there were the two kinds of errors.

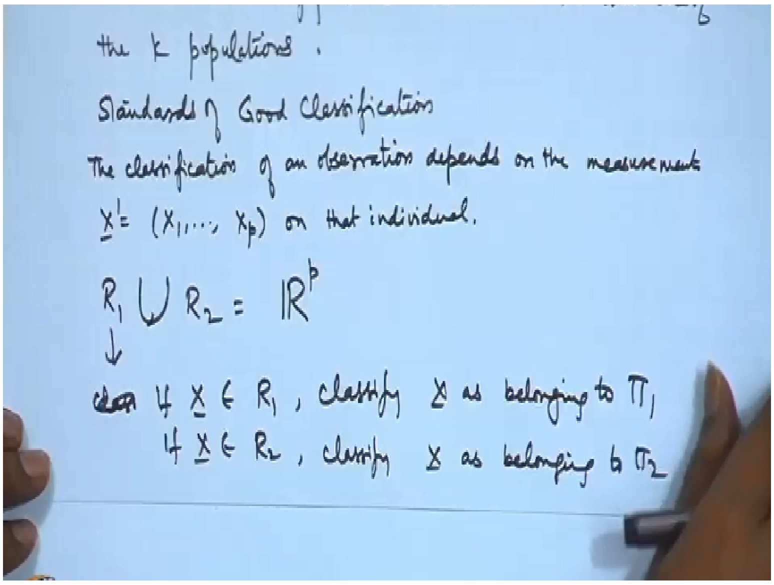
But when we are dealing with the k population here in the classification then the probability of misclassification or the probability of correct classification also becomes manifold that means an observation could have belonged to pi 1 and we classify it as pi 2 then observation could have been from pi 1 we could have classify it as pi 3 and so on and similarly the other way round that means the observation could be from any of the pi j.

And we can classify it as one of the pi i. Along with that we can also have the cost of misclassification along with the probability another additional thing could be that if you do the wrong classification then there can be some additional cost. So in a general decision theoretic setup one can also consider that. The particular case can be that if you have a correct classification you have no loss and if no cost is implemented.

And if you make a wrong classification then you are incurring say one cost then you can get a 0, 1 loss function kind of thing. So now let me give some notation here the classification of an observation depends on the measurements=x1, x2, xp on that individual. So we can actually consider R1 and R2 as a partition of the p dimensional space here where R 1 is the space where classify that is if x belongs to say R1 classify x as belonging to pi 1.

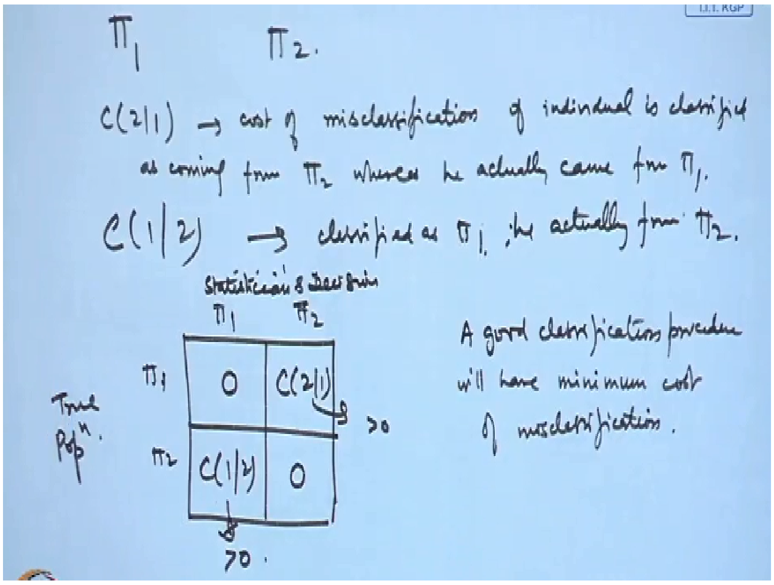
And if x belongs to R2.

(Refer Slide Time: 43:05)



Then you classify x as belonging to pi 2. This R1 and R2 are disjoint regions in the p dimensional sample space.

(Refer Slide Time: 43:25)



I mentioned about the kind of errors suppose in the beginning we consider only 2 populations say pi 1 and pi 2. So we may consider C/2/1 as the cost of misclassification if individual is classified as coming from pi 2 whereas he actually came from pi 1. Similarly, we can consider C/1/2 that is classified as pi 1 and he actually came from pi 2. So we have 2 cost of misclassification.

So in a decision theoretic setup if we consider it as a loss metric we can guide it in this fashion pi 1, pi 2 that is the statistician decision and on this side we have pi 1, pi 2 that is the true population. So if the true population is pi 1 we classify it as pi 1 then there is a 0 cost. Similarly, if true population is pi 2 and we classify as pi 2 then also it is 0. If the true population is pi 1 and we classify it as 2 then the cost is C1 2 1 and similarly here the cost is C/2/1.

So these two terms are taken to be positive in general. A good classification procedure will have minimum cost of misclassification. As we have seen in the previous discussion in general in the statistical decision making problem it is not possible to completely minimize the misclassification cost like in the case of testing of hypothesis problems also we have seen that the type 1 error and the type 2 error cannot be completely eliminated.

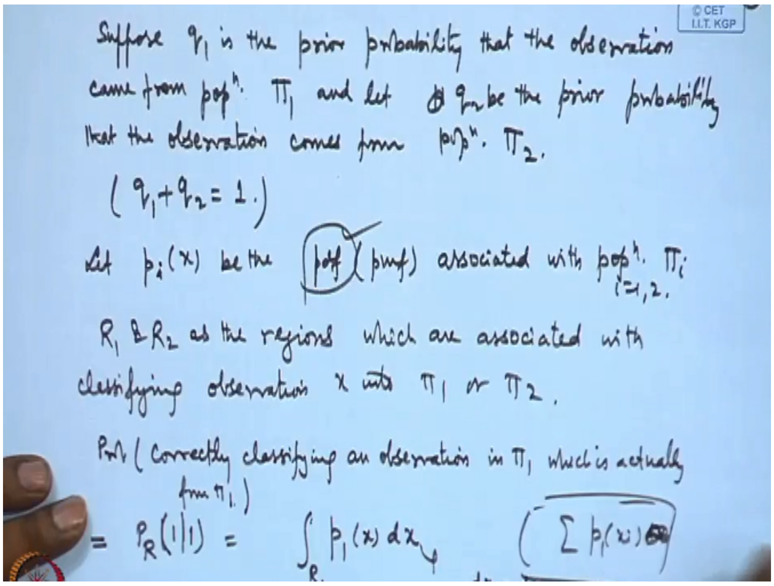
So there was a compromise which was worked out that you can consider fixed level of significance and then you consider the probability of type 2 error to be the smallest or the power of the test to be the maximum. So this was one of the compromise solutions that we considered. So if we consider it as a true population then it is actually a part of you can consider it as a testing of hypothesis problem.

And therefore both cannot be minimized simultaneously. So let us consider here this cost function and in what way we can consider the minimization etcetera. So one type of terminology which we did not consider in the testing of hypothesis problem is to allocate prior probabilities to each of the population. For example, if we know that both the population may occur with equal probabilities or the population 1 may occur with probability 1/3.

And population 2 may occur with probability 2/3 and so on. For example, you get a satellite image and you want to classify whether it is a land area or whether it is a water area. So if the image is taken from the satellite of the earth area a portion of the earth then you know that earth area is say the land area in the whole earth is 1/4 and the water area is 3/4. So you can allocate the probability p1 and P2 the prior probability.

So if you have the prior probabilities then we can reduce this number the probability of misclassification to a single number. So you can consider the Bayesian Classification Rules.

(Refer Slide Time: 48:18)

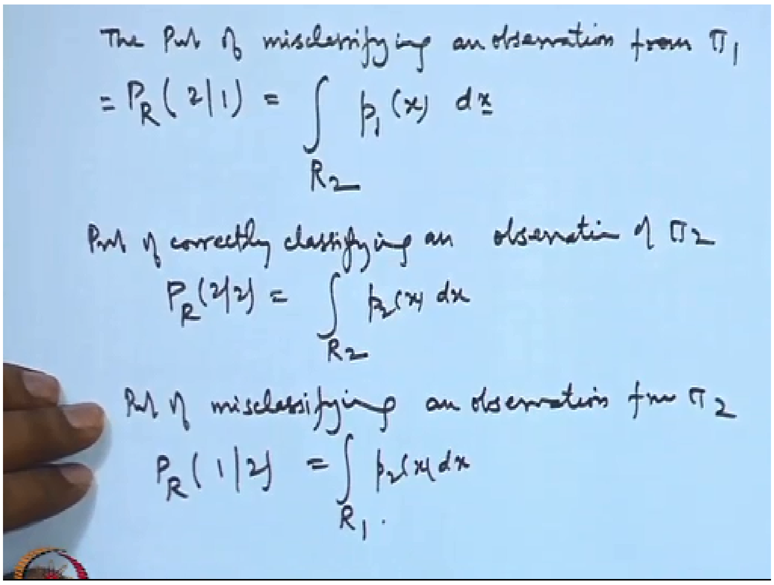


So let me introduce this here now. Suppose q1 is the prior probability that the observation came from population pi 1 and let q2 be the prior probability that the observation comes from population pi 2. So here of course q1+q2 should be=1. Let us consider say p i x is the so you may have a discrete or continuous distribution or it could be mixture also, but in particular let us take either purely discrete or purely continuous.

So you will have a pdf or pmf associated with the population pi i and we are considering R1 and R2 as the regions which are associated with classifying observation x into pi 1 or pi 2. So we define probability of correctly classifying an observation in pi 1 which is actually from pi 1. This we write as pR 1/1. This we can write as integral p1 x dx. So I am considering actually the density function form if it is a discrete case we can equivalently change it to the summation also.

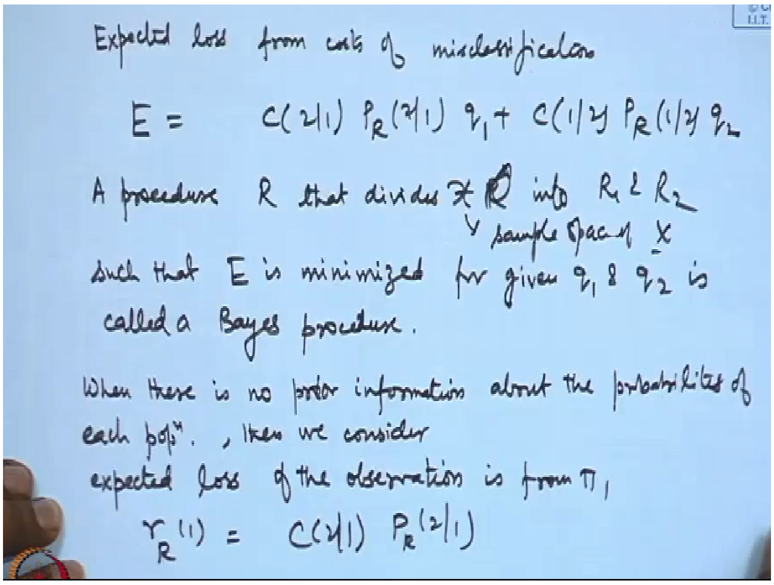
So I am not discussing this case separately let us have this interpretation. So this dx is actually it would be multivariate because it depends upon what kind of observation you are having. In general, we may be dealing with multivariate observations here.

(Refer Slide Time: 51:34)



So similarly we can define the probability of misclassifying an observation from pi 1 that is P R2 given 1 that means it is coming from 1 we classify it as 2 that means the density is actually p1, but we put it as R2. And in a similar way we have the probability of correctly classifying an observation of pi 2 that will be PR 2/2 and probability of misclassifying an observation from pi 2 that we will write as PR 1/2. That will be integral p2 x dx R1.

(Refer Slide Time: 53:08)



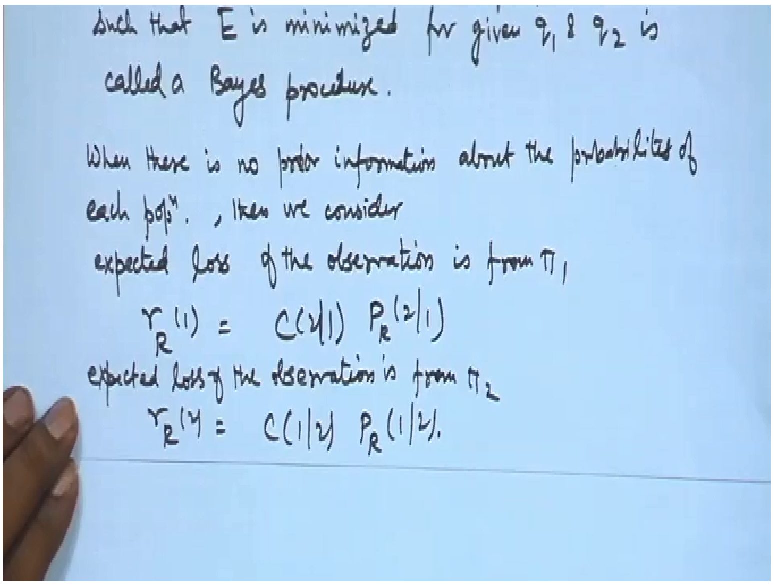
So we can consider now expected loss from cost of misclassification. Let us call it E that will be =C/2/1 PR 2/1 q1+C/1/2 PR 1/2 q2. Let me explain this if the observation is from 1 so the prior probability of the population 1 is q1 and we are incorrectly classifying it into 2. So the probability is this and we also incur a cost C/2/1 of misclassification. Similarly, if the population is actually 2 then the prior probability is q2.

And we misclassify it as 1 so the probability of that is PR 1/2 and then the cost of misclassifying an observation from 2 into 1 that is C/1/2. So this becomes the expected loss. So a procedure R that divides Rn into R1 and R2. So we consider this sample space say x not in Rn because I have not mentioned the dimension here. Let us consider x here this is the sample space of x into R1 and R2 such that E is minimized for given q1 and q2.

This is called a Bayes procedure. So we will mention that how to obtain a Bayes procedure here. There can be another way when there is no prior information then I will have 2 different terms that is the probability of misclassification from first one and the probability of misclassification on the second one. So let me define that also. When there is no prior information about the probabilities of each population.

Then we consider 2 terms. Expected loss if the observation is from pi 1 that is we call r R1 that is=C/2/1*PR 2 given 1.

(Refer Slide Time: 56:37)



And similarly expected loss if the observation is from pi 2 that we call rR2 that is=C/1/2 PR 1/2. We can give some decision theoretic definition which I will be explaining in the next lecture like we can call about a procedure being better than another procedure, a procedure being as good as another procedure and admissible procedure, a Minimax procedure and we will show that when the prior probabilities are known the Bayesian procedure can be determined.

When the prior probabilities are not known we will try to find out the Minimax procedure.

We will also develop the procedures for classification into multivariate normal populations. So these things we will be covering in the following lecture.