ANÁLISE DISCRIMINANTE

DISSERTAÇÃO APRESENTADA AO
INSTITUTO DE MATEMÁTICA E ESTATÍSTICA
DA
UNIVERSIDADE DE SÃO PAULO
PARA A OBTENÇÃO DO GRAU DE MESTRE
EM
ESTATÍSTICA APLICADA

ORIENTADOR:
PROF.DR. CLÓVIS DE ARAUJO PERES

- SÃO PAULO, AGOSTO DE 1975 -

PREFÁCIO

A Análise Discriminante é um conjunto de técnicas estatísticas cuja finalidade é a locação de um elemento em uma de k populações previamente conhecidas, suposto que esse elemento pertence realmente a uma delas.

No presente trabalho procuramos expor de maneira didática os principais aspectos da Análise Discriminante. O primeiro Capítulo se destina à apresentação da Análise Discriminante, das principais dificuldades que ocorrem nas aplicações práticas e de um resumo histórico do desenvolvimento desse assunto. No Capítulo 2 abordamos a função discriminante linear proposta por Fisher para a discriminação entre duas populações. No Capítulo 3 é feito um estudo geral da discriminação entre duas populações considerando-se as probabilidades de erro de classificação. O Capítulo 4 trata da discriminação entre k populações ($k \geq 2$). Finalmente, o Capítulo 5 é um exemplo de aplicação de técnicas desenvolvidas nos Capítulos anteriores.

* * *

Gostaríamos de externar os nossos agradecimentos aos professores Dr. Clóvis de Araujo Peres (orientador), Dr. Boris Schneiderman, Melvin Cymbalista e Paulo Brito Leme pelas valiosas colaborações e ao sr. João Baptista Esteves de Oliveira, que pacientemente datilografou os originais deste trabalho.

- □ -

I N D I C E

CAPÍTULO 1 - INTRODUÇÃO

1.1 - O que é Análise Discriminante 1

1.2 - Algumas dificuldades nas aplicações práticas da Análise Discriminante. 3

1.3 - A escolha conveniente de variáveis. 6

1.4 - Resumo histórico. 7

CAPÍTULO 2 - DISCRIMINAÇÃO ENTRE DUAS POPULAÇÕES - A FUNÇÃO DISCRIMINANTE DE FISHER

2.1 - Introdução. 15

2.2 - A função Discriminante. 15

2.3 - Obtenção dos coeficientes da função discriminante 17

2.4 - A discriminação do elemento E 19

2.5 - Variância da função discriminante 21

2.6 - Teste da significância da discriminação . . . 22

2.7 - Um exemplo numérico 24

2.8 - Interpretação geométrica. 27

2.9 - Um paralelo com regressão multilinear 29

CAPÍTULO 3 - DISCRIMINAÇÃO ENTRE DUAS POPULAÇÕES - OUTROS CRITÉRIOS

3.1 - Introdução. 33

3.2 - Determinação de R quando as probabilidades a priori são conhecidas 36

3.3 - Determinação de R quando as probabilidades a priori são desconhecidas. 37

3.4 - Aplicação para o caso de populações normais . 41

3.5 - Relacionamento da função discriminante de Fisher com a solução Bayesiana. 48

3.6 - A função discriminante quadrática 51

3.7 - A região de dúvida de Rao 52

3.8 - Outros critérios discriminantes 55

3.9 - Testes relacionados com a discriminação entre duas populações 55

CAPÍTULO 4 - DISCRIMINAÇÃO ENTRE k POPULAÇÕES

4.1 - Introdução. 63

4.2 - Determinação de R quando as probabilidades a priori são conhecidas 64

4.3 - Determinação de R quando as probabilidades a priori são desconhecidas. 66

4.4 - Discriminação entre k populações normais. . 68

4.5 - A região de dúvida. 73

4.6 - Dois testes relacionados com a discriminação entre k populações. 76

4.7 - A escolha conveniente de variáveis. 81

CAPÍTULO 5 - EXEMPLO DE APLICAÇÃO

5.1 - O problema. 84

5.2 - Apresentação dos dados. 84

5.3 - Teste da significância da discriminação . . 86

5.4 - A discriminação 87

5.5 - Probabilidade de erro de classificação. . . 90

5.6 - Uma região de dúvida. 91

5.7 - A escolha das variáveis discriminantes. . . 93

APÊNDICE A. 96

APÊNDICE B. 99

APÊNDICE C. 102

APÊNDICE D. 104

BIBLIOGRAFIA. 106

CAPÍTULO 1

INTRODUÇÃO

1.1 - O QUE É ANÁLISE DISCRIMINANTE

Entendemos a Análise Discriminante como um conjunto de processos estatísticos com a finalidade de locar um indivíduo E em uma de k populações distintas, previamente conhecidas, admitindo-se que E realmente pertença a uma dessas k populações. O tratamento estatístico dado a esse problema de locação reside no fato de que os dados utilizados são os valores de um conjunto de p variáveis aleatórias $(X_1, \ldots, X_p)$.

Por exemplo, um crânio é descoberto em escavações arqueológicas e deseja-se saber se pertenceu a um ser do sexo masculino ou feminino. Comparando p medidas do crânio encontrado com as respectivas medidas dos crânios de elementos do sexo masculino e do sexo feminino tentaremos discriminar o sexo do crânio. O objetivo na Análise Discriminante é tentar estabelecer a melhor regra de comparação possível, isto é, aquela que minimize o erro de discriminação.

Alguns autores consideram a Análise Discriminante

com objetivos mais amplos. Para eles a Análise Discriminante não engloba apenas um conjunto de técnicas com objetivos de decisão — a locação do elemento E — mas também consiste no estudo descritivo das k populações consideradas, ou seja estuda-se a "configuração" do conjunto das k populações. No entanto, no presente trabalho trataremos apenas da Análise Discriminante sob o ponto de vista de decisão.

As expressões "discriminar" e "classificar" tem, para muitos autores sentidos diferentes mas nós as confundiremos principalmente movidos pela sua identidade no linguajar do dia a dia. Rao (referência 10) sugere ainda o termo "identificar" como o mais apropriado.

Segundo Kendall (referência 7) as dificuldades que surgem ao classificar um indivíduo, dificuldades estas que geram a própria existência da Análise Discriminante, podem ser resumidas em:

- *informação perdida:* é o caso do crânio, descrito anteriormente, que deve ser classificado como masculino ou feminino. A informação principal está perdida e devemos recorrer a outros atributos que permitam a classificação.
- *informação inacessível:* se a necessidade de um determinado tipo de cirurgia não puder ser diagnosticada com certeza pela análise clínica do paciente, temos caracterizado um caso de informação inacessível: a informação existe mas não é acessível no sentido de que se possa diagnosti-

car a necessidade de cirurgia sem ter que recorrer à mesma. O que pode ser feito é escolher o melhor conjunto de análises clínicas disponíveis de forma a minimizar a probabilidade de errar o diagnóstico.

- *informação inexistente:* ocorre nos casos de predição. Um novo sistema de transportes vai ser implantado e deseja-se classificar um indivíduo como futuro usuário ou não. Baseados em experiências anteriores com sistemas de transportes equivalentes tentaremos fazer tal classificação, mas a certeza só virá após a entrada em funcionamento do novo sistema.

1.2 - ALGUMAS DIFICULDADES NAS APLICAÇÕES PRÁTICAS DA ANÁLISE DISCRIMINANTE

Para bem entender quais as dificuldades práticas na aplicação das técnicas de classificação é importante considerar o seguinte:

– A cada elemento das k populações consideradas associaremos um vetor $(x_1,\ldots,x_p)$ cujos componentes representam os valores de p variáveis aleatórias $X_1,\ldots,X_p$ para esse elemento. Estas variáveis são as medidas em que irá basear-se o processo discriminatório. Para o estabelecimento deste último devemos conhecer a distribuição conjunta das variáveis $X_1,\ldots,X_p$ em cada uma das k populações, como veremos mais adiante.

— O primeiro problema surge do desconhecimento, em muitos casos de aplicação, da função densidade das variáveis consideradas. Às vezes, conhecemos o tipo de distribuição mas ignoramos o valor dos parâmetros que a caracterizam. Nestes casos, a saída é a utilização das estimativas desses parâmetros. Isto em geral aumenta o erro de classificação especialmente nos casos em que as amostras não são suficientemente representativas.

— As amostras, de cada uma das k populações, a partir das quais se estimarão os parâmetros populacionais podem apresentar-se de duas formas distintas: - dispõe-se de k amostras bem definidas das quais se conhece a população de origem ou os elementos amostrais das k populações encontram-se misturados. No primeiro caso não surgem dificuldades mas no segundo o problema se complica e muito. Rao (referência 9) descreve um método para a separação de duas populações normais unidimensionais. Felizmente, quase sempre é possível contar com amostras já separadas.

— Ao classificar o elemento E é importante considerar as probabilidades *a priori* p_i, isto é, a probabilidade que um elemento escolhido ao acaso, dentre todos os elementos das k populações, pertença a população i. No entanto tais probabilidades a priori nem sempre são conhecidas e às vezes nem mesmo estimáveis a partir dos dados com que contamos.

Segundo recente artigo escrito por C.R. Rao (refe

rência 11) as aplicações da análise discriminante tem sido poucas devido especialmente à falta de informação sobre a distribuição das p variáveis consideradas e sobre as probabilidades a priori das k distribuições. E a dificuldade tem sido maior no caso de algumas variáveis terem distribuição discreta.

— Outro elemento que também deve ser considerado é o custo decorrente de uma classificação errada ou seja quando um elemento da população i é classificado como sendo da população j ($i \neq j$). Em muitas aplicações a avaliação de tais custos é bastante difícil. O conhecimento das relações entre tais custos resolve em parte o problema. No entanto, quando nada se sabe a respeito dos custos é comum considerá-los iguais.

— Nos processos discriminatórios que apresentaremos, em muitas ocasiões (especificamente no caso de populações normais) consideraremos que as diferentes populações apresentadas têm mesma dispersão. Existem testes para verificar a homogeneidade das populações (Wilks, Roy). Mas como estes testes são de aplicação bastante difícil é usual assumir a igualdade das dispersões, visto que pequenos desvios pouco afetarão o efeito discriminante dos processos em questão.

— Analisaremos agora uma dificuldade prática diferente das acima descritas: o elemento a ser classificado é por exemplo um ser do qual apenas temos o crânio encontrado em esca

vações arqueológicas. Se o crânio deve ser classificado como de um ser masculino ou feminino não há dúvida que a única coisa a fazer é comparar as medidas do crânio com as dos dois sexos de modo a identificá-lo. Com certeza se não for masculino será feminino ou vice-versa, isto é, a existência de apenas dois grupos é certa e pode ser considerada um dado do problema. No entanto se o ser a que pertenceu o crânio deve ser classificado como proveniente de um povo A ou de um povo B, não é correto aceitar que a pertinência a um destes grupos seja um dado do problema. Devemos admitir a possibilidade que o ser possa ter pertencido a um terceiro povo C. Assim um primeiro passo, que deve preceder a discriminação, é verificar a pertinência do elemento a ser classificado aos grupos A e B. Rao (referência 10) descreve o procedimento a seguir para tal verificação no caso de suas populações e indica a extensão do mesmo para mais de 2 populações.

— As dificuldades apresentadas até aqui são as que nos parecem as mais comuns. Evidentemente outras dificuldades existem nos diferentes casos de aplicação e inerentes aos mesmos. Não caberia portanto citá-las aqui.

1.3 - A ESCOLHA CONVENIENTE DE VARIÁVEIS

Um dos problemas da Análise Discriminante reside na escolha conveniente das variáveis X_i sobre as quais se ba-

seará o processo discriminatório. Importante, especialmente do ponto de vista de custo operacional, é selecionar dentre as p variáveis medidas o menor conjunto de q variáveis, tal que a discriminação baseada nessas q variáveis não só seja significativa,como também não possa ser "melhorada" significativamente com a inclusão das p-q variáveis restantes.

Poderíamos, a primeira vista, considerar todos os subconjuntos possíveis dessas p variáveis (2^p-1 subconjuntos) e compará-los quanto à discriminação. Mas isto seria um trabalho absurdo especialmente se p não for pequeno (para p = 10 teremos 1.023 subconjuntos!).

Normalmente adota-se um processo stepwise (passo a passo) que basicamente consiste do seguinte: dentre as p variáveis iremos selecionando sucessivamente (passo a passo) a melhor variável discriminadora, as duas melhores variáveis discriminadoras,..., até que nenhuma das novas variáveis a serem selecionadas introduza uma melhoria significativa na discriminação. Um processo stepwise muito utilizado é o BMD07M (do Biomedical Computer Programs), que será abordado mais adiante.

1.4 - RESUMO HISTÓRICO

Apresentaremos neste ítem um resumo histórico do desenvolvimento da Análise Discriminante do ponto de vista classificatório, mostrando algumas das principais linhas na

procura das soluções.

– A primeira solução para o problema de discriminação entre (duas) populações pode ser considerada a *função discriminante* proposta por Fisher em 1935. Seu primeiro artigo sobre o assunto, "The use of multiple measurments in taxonomic problems" foi publicado em 1936 e se dedicava à discriminação entre dois tipos de plantas.

Sem justificação teórica, Fisher sugeriu o uso de uma função linear das variáveis $X_1,\ldots,X_p$, que caracterizam os elementos das duas populações, cujos coeficientes deviam ser calculados de forma que a função maximizasse a "distância entre populações" definida pelo quociente entre a diferença das médias amostrais da função linear e o desvio padrão da mesma dentro (within) das amostras.

A obtenção exata dos coeficientes b_i era bastante laboriosa, pois exige a inversão de uma matriz de ordem p, o que gerou uma série de trabalhos sugerindo aproximações no cálculo da função discriminante. Os trabalhos de Beall e Jackson são os mais representativos nessa área.

Apesar de que a função discriminante foi sugerida para o problema específico de discriminação entre duas populações seu uso foi estendido desde o início (Barnard, 1935) para a discriminação entre mais de duas populações. No entanto sendo um escalar, a função discriminante não esgota a informação a respeito da configuração (disposição geométri-

ca) de mais de duas populações a menos que estas sejam colineares (os centros das populações estão sobre uma reta). É claro que o conhecimento da configuração das populações é importante no problema de discriminação. Ciente disto, Fisher sugeriu em 1938 um teste para verificar a colinearidade de k populações. Um estudo sobre a extensão da função discriminante para k populações foi apresentado por Bryan (1951) com a introdução das "funções discriminantes múltiplas". Estas diferentes funções discriminantes que originalmente foram construídas para classificar um elemento em uma das k populações, tem vital importância no estudo das direções em que variam as populações (configuração das populações).

Durante muito tempo a Análise Discriminante se resumiu à aplicação da função discriminante quer na forma original apresentada por Fisher, quer com aproximações ou extensões.

Mas tarde, ficou provado que sob o ponto de vista de minimizar as probabilidades de erro de classificação, a função discriminante proposta por Fisher é a solução indicada para o caso de duas populações normais com mesma matriz de covariância.

— O enfoque dado à discriminação sob o ponto de vista da função discriminante de Fisher é naturalmente baseado em razões intuitivas (maximizar a distância entre populações) não

levando em consideração o problema de minimização das probabilidades de erro no processo discriminatório.

As noções de erro na aplicação de técnicas estatísticas foram introduzidas por Neyman e Pearson. Welch (1939) em um artigo "Note on Discriminant Functions" abordou o processo de discriminação entre duas populações adotando as idéias de Neyman e Pearson, isto é, considerando as probabilidades de erro de classificação.

Basicamente a solução proposta por Welch consiste em particionar o espaço amostral em duas regiões R_1 e R_2 de forma tal que se um elemento pertencer a R_1 será classificado na população 1 e caso contrário na população 2. A escolha da partição deve ser feita de modo a minimizar a probabilidade de erro de classificação. Dois casos distintos foram considerados: as probabilidades a priori conhecidas e as probabilidades a priori desconhecidas. Em ambos os casos ficou demonstrado que as partições são definidas com auxílio da razão de verossimilhança.

— Uma das desvantagens do trabalho de Welch era a restrição ao caso de duas populações. Em 1945 Von Mises apresentou um trabalho em que estudava a discriminação entre k ($k \geq 2$) populações. Von Mises abordou o problema de uma forma um pouco distinta de Welch, trabalhando com as probabilidades de classificação correta associadas a cada população.

— Mais trade, Rao indtroduz um novo conceito nos problemas

de discriminação. Até então o espaço amostral na discriminação entre k populações era particionado em k regiões $R_1,\ldots,R_k$. A modificação sugerida por Rao foi a partição em k+1 regiões R_0, R_1, ..., R_k em que R_0 foi denominada região de dúvida e utilizada da seguinte forma: se o elemento a classificar pertencer a R_0 nenhuma decisão será tomada quanto à sua classificação. O aparecimento de R_0 é motivado pela construção de $R_1,\ldots,R_k$ sob limites pré-estabelecidos das probabilidades de classificação errada. Rao estendeu para k populações a solução proposta por Welch quando as probabilidades a priori são conhecidas. No caso destas últimas não serem conhecidas sugere que sejam consideradas iguais.

— O inconveniente destes processos discriminatórios construídos sob a consideração das probabilidades de erro é que assumem como conhecidas as distribuições conjuntas das p variáveis $X_1,\ldots,X_p$ que caracterizam os elementos das populações. Como já dissemos anteriormente estas distribuições nem sempre são conhecidas e a saída óbvia é a substituição nas regras discriminantes dos parâmetros populacionais por suas estimativas. Alguns estatísticos ocuparam-se do estudo das distribuições amostrais das estatísticas utilizadas nas regras discriminatórias com a finalidade de estudar as probabilidades de erro, e nem sempre tais distribuições são de fácil utilização. Wald estudou a distribuição amostral da função discriminante proposta por Fisher chegando a resulta

dos bastante complicados.

— Em 1949 um artigo publicado por Hoel e Peterson abre uma nova linha de estudos na Análise Discriminante. Neste trabalho a estimação dos parâmetros populacionais, das probabilidades a priori e a discriminação são estudados simultaneamente. A idéia básica é a obtenção de um conjunto de estimadores que maximize a probabilidade de classificação correta.

— Uma nova linha de pesquisas foi iniciada na Análise Discriminante devido a um problema proposto por Burnaby em 1966. Os problemas que vinham sendo tratados até então resumiam-se às classificações de elementos em uma de k populações. No caso do problema proposto por Burnaby trata-se de classificar um elemento em um de dois conjuntos, cada um dos quais constituído por uma mescla de diferentes populações em proporções desconhecidas. Rao abordou o assunto propondo ainda em 1966 uma solução para o mesmo.

— Muitos assuntos continuam merecendo um maior estudo dentro da Análise Discriminante, por exemplo: métodos para escolha das variáveis, discriminação entre populações com mesmas médias porém diferentes dispersões. E se o aperfeiçoamento depende da necessidade, não cremos que pare o aprimoramento na Análise Discriminante.

— Referências bibliográficas dos trabalhos citados no resumo histórico:

[1] - BERNARD, M.M. - 1935 - "The secular variations of

skull characteres in four series of egiptian skulls" - Annals of Eugenics, 6.

[2] - BEALL, G - 1945 - "Approximate methods in calculating discriminant functions" - Psychometric, 10.

[3] - BRYAN, J.G. - 1951 - "The generalized discriminant function: Mathematical foundation and computational routine" - Havard Educational Review, 21.

[4] - BURNABY, T.P. - 1966 - "Growth invariant discriminant functions and generalized distances" -Biometrics, 22.

[5] - FISHER, R.A. - 1936 - vide Bibliografia.

[6] - FISHER, R.A. - 1938 - vide Bibliografia.

[7] - HOEL, P.G e PETERSON, R.P. - 1949 - "A solution to the problem of optimum classification"-Annals of Mathematical Statistics, 20.

[8] - JACKSON, R.W.B. - 1943 - "Approximate multiple regression weights" - Journal of Experimental Education, 11.

[9] - RAO, C.R. - 1948 - "The utilization of multiple measurments in problems of biological classification" - Journal of the Royal Statistical Society, Series B, 10.

[10] - RAO, C.R. - 1966 - "Discriminant function between composite hypotheses and related problems" - Biometrika, 53.

[11] - VON MISES, R. - 1945 - " On the classification of observation data into distinct groups" - Annals of Mathematical Statistics, 16.

[12] - WALD, A. - 1944 - "On a statistical problem arising in the classification of an individual into one of two groups" - Annals of Math. Statistics, 15.

[13] - WELCH, B.L. - 1939 - "Note on discriminant functions" Biometrika, 31.

- □ -

CAPÍTULO 2

DISCRIMINAÇÃO ENTRE DUAS POPULAÇÕES

A FUNÇÃO DISCRIMINANTE DE FISHER

2.1 - INTRODUÇÃO

Embora a função discriminante de Fisher não seja sempre a regra mais apropriada para resolver o problema de discriminação entre duas populações, as razões abaixo expostas nos levaram a dedicar um capítulo a essa função:

I - o valor histórico da função discriminante como primeira solução específica para o problema da discriminação,

II - a Análise Discriminante durante muito tempo se resumiu ao uso dessa função,

III - no caso muito comum de discriminação entre 2 populações normais de mesma matriz de covariância, a função discriminante apresenta algumas propriedades ótimas.

2.2 - A FUNÇÃO DISCRIMINANTE

Como sabemos, o problema é o estabelecimento de uma regra que permita da melhor maneira possível classificar

um elemento E em uma de duas populações π_1 e π_2, admitindo-se que E pertença a uma delas. Tal processo basea-se no conjunto de valores do vetor aleatório $(X_1,\ldots,X_p)$ associado a cada elemento de π_1 e π_2. Admitiremos conhecer uma amostra de π_1 com n_1 elementos e outra de π_2 com n_2 elementos.

O que Fisher propos em seu artigo (ref.3) para solução do problema foi a obtenção de uma função linear Y

$$Y = b_1X_1 + b_2X_2 + \ldots + b_pX_p$$

que maximize o quociente entre a diferença dos valores de Y calculados para as médias amostrais e o desvio padrão de Y estimado "dentro" das amostras.

Em outra palavras o problema se resume à otimização dos coeficientes $b_1,\ldots,b_p$ de forma que a função linear Y assim obtida seja a melhor discriminante entre π_1 e π_2 de todas as funções lineares Y, entendendo-se por melhor discriminação a maximização do quociente acima descrito.

É interessante observar que Fisher não justificou o porque da escolha de uma função linear e nem o porque do quociente a ser maximizado. Cremos que as razões intuitivas que o levaram ao estabelecimento desse quociente podem ser interpretadas como a obtenção de uma função que maximize a distância entre as duas populações, distância essa padronizada em termos de desvio padrão. Além disso deve-se observar que está implícito na definição do quociente que a dispersão das duas populações π_1 e π_2 é considerada igual.

2.3 - OBTENÇÃO DOS COEFICIENTES DA FUNÇÃO DISCRIMINANTE

Denotaremos por x_{ijk} o valor da variável X_j associado ao k-ésimo elemento da amostra da população π_i. Temos portanto $i=1,2$; $j=1,2,\ldots,p$; $k=1,2,\ldots,n_i$. Denotaremos por $(x_1,\ldots,x_p)$ os valores associados ao elemento E.

Sejam:

$$y_{ik} = b_1\, x_{i1k} + \ldots + b_p\, x_{ipk}$$

$$\bar{y}_i = \frac{1}{n_i} \sum_i y_{ik} \qquad \text{onde } \sum_i = \sum_{k=1}^{n_i}$$

$$d_j = \frac{1}{n_1} \sum_1 x_{ijk} - \frac{1}{n_2} \sum_2 x_{2jk} = \bar{x}_{1j} - \bar{x}_{2j}$$

Maximizar o quociente proposto por Fisher é equivalente a maximizar o quociente

$$\frac{(\bar{y}_1-\bar{y}_2)^2}{\sum_1 (y_{1k}-\bar{y}_1)^2 + \sum_2 (y_{2k}-\bar{y}_2)^2} = \frac{d^2}{S} \qquad (2.3.1)$$

onde S é proporcional à variância amostral "dentro" de Y.

Como

$$d = \bar{y}_1-\bar{y}_2 = b_1(\bar{x}_{11}-\bar{x}_{21})+\ldots+b_p(\bar{x}_{1p}-\bar{x}_{2p}) = \sum_{j=1}^{p} b_j d_j$$

$$\sum_1 (y_{1k}-\bar{y}_1)^2 + \sum_2 (y_{2k}-\bar{y}_2)^2 =$$

$$= \sum_{i=1}^{2} \sum_{k=1}^{n_i} \left\{ \sum_{j=1}^{p} \sum_{m=1}^{p} b_j b_m (x_{ijk}-\bar{x}_{ij})(x_{imk}-\bar{x}_{im}) \right\} = \sum_{j=1}^{p} \sum_{m=1}^{p} b_j b_m S_{jm}$$

onde $S_{jm} = \sum_{i=1}^{2} \sum_{k=1}^{n_i} (x_{ijk}-\bar{x}_{ij})(x_{imk}-\bar{x}_{im})$

podemos reescrever (2.3.1) da seguinte forma

$$\frac{(\sum_{j=1}^{p} b_j d_j)^2}{\sum_{j=1}^{p} \sum_{m=1}^{p} b_j b_m S_{jm}} = \frac{d^2}{S}$$

onde observamos que o conjunto de coeficientes $b_1,\ldots,b_p$ que maximiza o quociente não é único, porém dois quaisquer dos possíveis conjuntos de soluções diferirão entre si apenas por um fator constante. Isto significa que entre todas as soluções possíveis permanecem constantes (p-1) relações, a saber $\lambda_1 = (b_1/b_2)$, $\lambda_2 = (b_2/b_3),\ldots,$ $\lambda_{p-1} = (b_{p-1}/b_p)$. Desta forma no processo de maximização podemos ignorar constantes que multipliquem todos os coeficientes.

Maximizando o quociente, teremos para cada b_j:

$$\frac{\partial}{\partial b_j}(\frac{d^2}{S}) = 0 \Longrightarrow \frac{d}{S}(2\frac{\partial d}{\partial b_j} - \frac{d}{S}\frac{\partial S}{\partial b_j}) = 0$$

$$S = \sum_{j=1}^{p} \sum_{m=1}^{p} b_j b_m S_{jm} \Longrightarrow \frac{\partial S}{\partial b_j} = 2\sum_{m=1}^{p} b_m S_{jm}$$

$$d = \sum_{j=1}^{p} b_j d_j \Longrightarrow \frac{\partial d}{\partial b_j} = d_j$$

resultando

$$d_j = \frac{d}{S}\sum_{m=1}^{p} b_m S_{jm}$$

Como (d/S) multiplica todos os coeficientes podemos ignorá-lo e portanto, um conjunto de coeficientes que maximiza o quociente proposto por Fisher será o conjunto de soluções do sistema

$$\sum_{m=1}^{p} b_m S_{jm} = d_j$$

ou

$$\begin{array}{l} b_1 S_{11} + b_2 S_{12} + \ldots + b_p S_{1p} = d_1 \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ b_1 S_{p1} + b_2 S_{p2} + \ldots + b_p S_{pp} = d_p \end{array}$$

e a resolução deste sistema vai se tornando cada vez mais trabalhosa a medida que p cresce, o que motivou uma série de artigos propondo aproximações do modelo de Fisher com a finalidade de simplificar os cálculos. A referência 6 da bibliografia mostra algumas das proposições de simplificação.

Note-se que o valor máximo do quociente (2.3.1) é numericamente igual a d, pois

$$S = \sum_{j=1}^{p} \sum_{m=1}^{p} b_j b_m S_{jm} = \sum_{j=1}^{p} b_j \sum_{m=1}^{p} b_m S_{jm} = \sum_{j=1}^{p} b_j d_j = \bar{y}_1 - \bar{y}_2 = d$$

onde d é não negativo pois S é uma soma de quadrados.

2.4 - A DISCRIMINAÇÃO DO ELEMENTO E

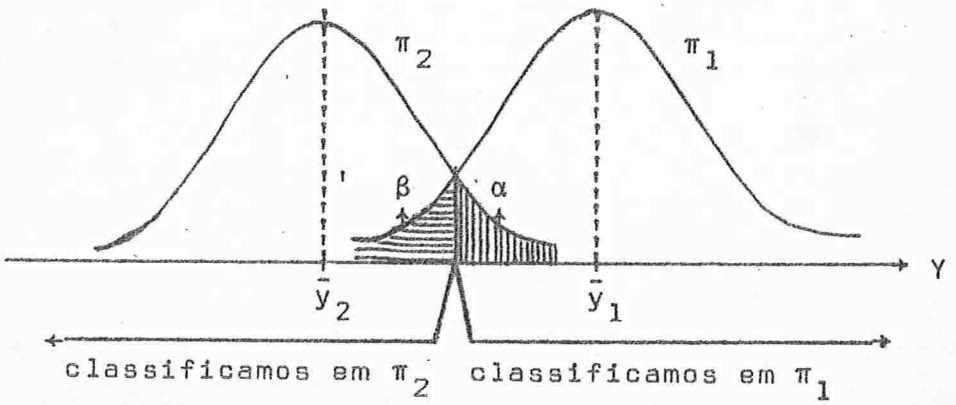
Indiquemos agora a utilização, proposta por Fisher, da função discriminante para classificar o elemento E.

Podemos calcular o valor da função discriminante para os vetores média $(\bar{x}_{i1},\ldots,\bar{x}_{ip})$ de cada amostra obtendo assim $\bar{y}_1$ e $\bar{y}_2$. O valor da função discriminante para o elemento E será $y=b_1x_1+\ldots+b_px_p$.

Se o valor de y for mais próximo de $\bar{y}_1$ do que de $\bar{y}_2$ classificaremos o elemento E na população π_1; caso contrário na população π_2. Se por acaso o valor de y coincidir com o valor médio de $\bar{y}_1$ e $\bar{y}_2$ classificaremos o elemento em π_1 ou π_2 indiferentemente.

Vemos portanto que o ponto médio entre $\bar{y}_1$ e $\bar{y}_2$ funciona como ponto de separação entre as duas populações. Devemos esclarecer no entanto que a escolha desse ponto como separação nem sempre é a melhor no sentido de minimizar a probabilidade de erro de classificação. No próximo capítulo veremos porque.

Supondo, por exemplo, que π_1 e π_2 são populações p-normais com mesma matriz de covariância (caso em que, veremos, a função discriminante de Fisher é a regra discriminante adequada) teremos esquematicamente



onde

α = probabilidade de classificar em π_1 um elemento que pertence a π_2

β = probabilidade de classificar em π_2 um elemento que pertence a π_1.

2.5 - VARIÂNCIA DA FUNÇÃO DISCRIMINANTE

Uma estimativa, w(y), da variância da função discriminante Y pode ser obtida da seguinte forma:

$$w(y) = \frac{\sum_{i=1}^{2} \sum_{k=1}^{n_i} (y_{ik}-\bar{y}_i)^2}{n_1+n_2-p-1} = \frac{S}{n_1+n_2-p-1}$$

O número de graus de liberdade

$$n_1+n_2-p-1 = n_1+n_2-2-(p-1)$$

é assim explicado por Fisher: 2 graus de liberdade são subtraídos justificados pelo uso das duas estimativas $\bar{y}_1$ e $\bar{y}_2$. Os outros (p-1) graus de liberdade subtraídos se justificam pela utilização dos coeficientes $b_1,\ldots,b_p$ estimados a partir das amostras e que estão vinculados por (p-1) relações constantes, a saber, $b_1/b_2,\ldots,b_{p-1}/b_p$.

De posse desta estimativa, conhecida a distribuição, podemos estimar a probabilidade de erro de classificação ao usar como regra discriminatória a função discriminante.

Alguns autores consideram o número de graus de liberdade simplesmente igual a n_1+n_2-2. Não concordamos com isto, pois ao adotar este número de graus de liberdade estamos implicitamente admitindo que os coeficientes b_i sejam constantes conhecidas, quando na verdade são estimados a partir das amostras consideradas.

2.6 - TESTE DA SIGNIFICÂNCIA DA DISCRIMINAÇÃO

Partindo da hipótese de que ambas as populações sejam normais com mesma matriz de dispersão podemos testar a significância da discriminação através de uma análise de variância, conforme sugestão de Fisher.

Estaremos testanto se a discriminação é ilusória no sentido de que as populações originais coincidam (em termos das variáveis $X_1,\ldots,X_p$) contra a hipótese de boa discriminação (não coincidência).

A soma de quadrados entre, SQE, será dada por:

$$SQE = \sum_{i=1}^{2} n_i(\bar{y}_i-\bar{\bar{y}})^2 \text{ onde } \bar{\bar{y}} = \frac{n_1\bar{y}_1+n_2y_2}{n_1+n_2}$$

ou

$$SQE = \frac{n_1\ n_2}{n_1+n_2}(\bar{y}_1-\bar{y}_2)^2 = \lambda^2 d^2 \text{ onde } \lambda^2 = \frac{n_1\ n_2}{n_1+n_2}$$

A soma dos quadrados dentro, SQD, já foi calculada anteriormente e é dada por:

$$SQD = S = \sum_{i=1}^{2} \sum_{k=1}^{n_i} (y_{ik}-\bar{y}_i)^2 = \bar{y}_1-\bar{y}_2 = d.$$

O quadro de análise resultante será dado por:

FONTE	G.L.	S.Q.	Q.M.
ENTRE AMOSTRAS	p	$\lambda^2 d^2$	$(\lambda^2 d^2)/p$
DENTRO DE AMOSTRAS	n_1+n_2-p-1	d	$d/(n_1+n_2-p-1)$
TOTAL	n_1+n_2-1	$d(1+\lambda^2 d)$	---

Mostremos agora que a estatística apropriada para o teste é uma F (de Snedecor) com p e n_1+n_2-p-1 graus de liberdade. O quociente entre quadrados médios será:

$$\frac{S_E}{S_D} = \frac{n_1+n_2-p-1}{p} \lambda^2 d.$$

Mas

$$\sum_{j=1}^{p} S_{ij} b_j = d_i \Longrightarrow b_i = \sum_{j=1}^{p} S^{ij} d_j \text{ onde } (S_{ij})^{-1} = (S^{ij})$$

portanto

$$d = \sum_{i=1}^{p} b_i d_i = \sum_{i=1}^{p} \sum_{j=1}^{p} S^{ij} d_i d_j = (n_1+n_2-2)^{-1} \sum_{i=1}^{p} \sum_{j=1}^{p} w^{ij} d_i d_j$$

onde $(w_{ij})^{-1} = (w^{ij})$ sendo $w_{ij} = \dfrac{S_{ij}}{n_1+n_2-2}$.

Como

$\sum_{i=1}^{p} \sum_{j=1}^{p} w^{ij} d_i d_j = D^2$, D^2: distância generalizada de Mahalanobis.

temos, conforme o Apêndice A

$$\frac{S_E}{S_D} = \frac{(n_1+n_2-p-1)}{(n_1+n_2-2)p}\lambda^2 D^2 \sim F_{p,n_1+n_2-p-1}.$$

É interessante observar que em termos de F podemos pensar na função discriminante como sendo aquela que maximiza a $F = S_E/S_D$.

2.7 - UM EXEMPLO NUMÉRICO

Para ilustrar a teoria exposta nos itens anteriores daremos agora um exemplo numérico. Este basear-se-á nos dados apresentados por Fisher em seu artigo "The use of multiple measurments in taxonomic problems" (referência 3).

Fisher apresentou a discriminação entre dois tipos de Iris (planta), a saber Iris Versicolor e Iris Setosa, baseada em 4 medidas de cada planta: comprimento do caule, largura do caule, comprimento da pétala, largura da pétala. Para tal, considerou uma amostra de cada tipo, ambas com 50 elementos.

O exemplo de Fisher é um tanto infeliz no seguinte sentido: para a Iris Versicolor o comprimento da pétala varia de 3,0 a 5,1 (na amostra) enquanto esta mesma medida varia de 1,0 a 1,9 para a Iris Setosa, o que torna fácil a discriminação baseada nessa única característica sem ter que recorrer a um processo estatístico de discriminação mais so

DADOS DO EXEMPLO DO ÍTEM 2.7

Iris versicolor				*Iris setosa*			
Sepal length	Sepal width	Petal length	Petal width	Sepal length	Sepal width	Petal length	Petal width
7·0	3·2	4·7	1·4	5·1	3·5	1·4	0·2
6·4	3·2	4·5	1·5	4·9	3·0	1·4	0·2
6·9	3·1	4·9	1·5	4·7	3·2	1·3	0·2
5·5	2·3	4·0	1·3	4·6	3·1	1·5	0·2
6·5	2·8	4·6	1·5	5·0	3·6	1·4	0·2
5·7	2·8	4·5	1·3	5·4	3·9	1·7	0·4
6·3	3·3	4·7	1·6	4·6	3·4	1·4	0·3
4·9	2·4	3·3	1·0	5·0	3·4	1·5	0·2
6·6	2·9	4·6	1·3	4·4	2·9	1·4	0·2
5·2	2·7	3·9	1·4	4·9	3·1	1·5	0·1
5·0	2·0	3·5	1·0	5·4	3·7	1·5	0·2
5·9	3·0	4·2	1·5	4·8	3·4	1·6	0·2
6·0	2·2	4·0	1·0	4·8	3·0	1·4	0·1
6·1	2·9	4·7	1·4	4·3	3·0	1·1	0·1
5·6	2·9	3·6	1·3	5·8	4·0	1·2	0·2
6·7	3·1	4·4	1·4	5·7	4·4	1·5	0·4
5·6	3·0	4·5	1·5	5·4	3·9	1·3	0·4
5·8	2·7	4·1	1·0	5·1	3·5	1·4	0·3
6·2	2·2	4·5	1·5	5·7	3·8	1·7	0·3
5·6	2·5	3·9	1·1	5·1	3·8	1·5	0·3
5·9	3·2	4·8	1·8	5·4	3·4	1·7	0·2
6·1	2·8	4·0	1·3	5·1	3·7	1·5	0·4
6·3	2·5	4·9	1·5	4·6	3·6	1·0	0·2
6·1	2·8	4·7	1·2	5·1	3·3	1·7	0·5
6·4	2·9	4·3	1·3	4·8	3·4	1·9	0·2
6·6	3·0	4·4	1·4	5·0	3·0	1·6	0·2
6·8	2·8	4·8	1·4	5·0	3·4	1·6	0·4
6·7	3·0	5·0	1·7	5·2	3·5	1·5	0·2
6·0	2·9	4·5	1·5	5·2	3·4	1·4	0·2
5·7	2·6	3·5	1·0	4·7	3·2	1·6	0·2
5·5	2·4	3·8	1·1	4·8	3·1	1·6	0·2
5·5	2·4	3·7	1·0	5·4	3·4	1·5	0·4
5·8	2·7	3·9	1·2	5·2	4·1	1·5	0·1
6·0	2·7	5·1	1·6	5·5	4·2	1·4	0·2
5·4	3·0	4·5	1·5	4·9	3·1	1·5	0·2
6·0	3·4	4·5	1·6	5·0	3·2	1·2	0·2
6·7	3·1	4·7	1·5	5·5	3·5	1·3	0·2
6·3	2·3	4·4	1·3	4·9	3·6	1·4	0·1
5·6	3·0	4·1	1·3	4·4	3·0	1·3	0·2
5·5	2·5	4·0	1·3	5·1	3·4	1·5	0·2
5·5	2·6	4·4	1·2	5·0	3·5	1·3	0·3
6·1	3·0	4·6	1·4	4·5	2·3	1·3	0·3
5·8	2·6	4·0	1·2	4·4	3·2	1·3	0·2
5·0	2·3	3·3	1·0	5·0	3·5	1·6	0·6
5·6	2·7	4·2	1·3	5·1	3·8	1·9	0·4
5·7	3·0	4·2	1·2	4·8	3·0	1·4	0·3
5·7	2·9	4·2	1·3	5·1	3·8	1·6	0·2
6·2	2·9	4·3	1·3	4·6	3·2	1·4	0·2
5·1	2·5	3·0	1·1	5·3	3·7	1·5	0·2
5·7	2·8	4·1	1·3	5·0	3·3	1·4	0·2

fisticado. O mesmo acontece com as larguras das pétalas.

No entanto se só considerarmos o comprimento e a largura dos caules estaremos diante de um problema que realmente requer a aplicação de um critério mais sotisficado para à discriminação. Utilizaremos neste caso a função discriminante.

Consideremos como população π_1 a correspondente à Iris Versicolor o como π_2 a correspondente a Iris Setosa. A variável X_1 corresponderá ao comprimento do caule e X_2 à largura do caule. A partir dos dados da tabela anexa obteremos

	I.VERSICOLOR	I.SETOSA	DIFERENÇAS
COMPRIMENTO DO CAULE	$\bar{x}_{11} = 5,936$	$\bar{x}_{21} = 5,006$	$d_1 = 0,930$
LARGURA DO CAULE . .	$\bar{x}_{12} = 2,770$	$\bar{x}_{22} = 3,428$	$d_2 = -0,658$

$$(S_{ij}) = \begin{pmatrix} 19,1434 & 9,0356 \\ 9,0356 & 11,8658 \end{pmatrix}$$

A obtenção dos coeficientes da função discriminante será através do sistema

$$19,1434b_1 + 9,0356b_2 = 0,930$$
$$9,0356b_1 + 11,8658b_2 = -0,658$$

que fornecerá

$$b_1 = 0,1167 \quad e \quad b_2 = -0,1443$$

e assim teremos a função discriminante

$$y = 0,1167x_1 - 0,1443x_2.$$

O valor de $\bar{y}_1$ média da função discriminante na amostra de π_1, será:

$$\bar{y}_1 = b_1\bar{x}_{11} + b_2\bar{x}_{12} = 0,2930$$

e o de $\bar{y}_2$ será:

$$\bar{y}_2 = b_1\bar{x}_{21} + b_2\bar{x}_{22} = 0,0895.$$

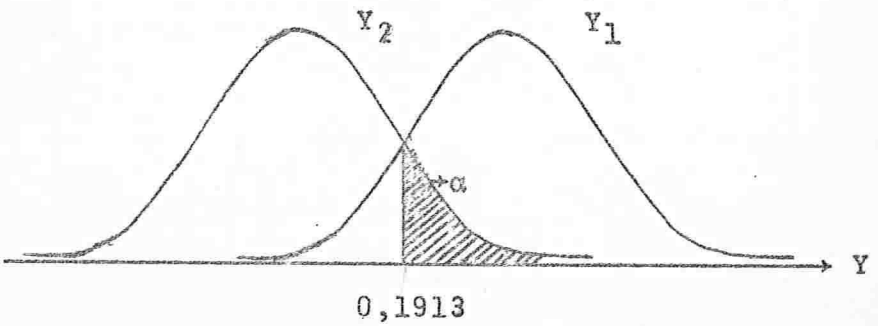
O ponto médio entre $\bar{y}_1$ e $\bar{y}_2$ é 0,1913. Assim, se uma nova planta Iris tiver que ser classificada como Versicolor ou Setosa (admitindo que se enquadre em um destes tipos) calcularemos a função discriminante correspondente e a classificaremos como Versicolor se esse valor for superior a 0,1913 ou como Setosa no caso contrário.

* * *

A estimativa w(y) da variância da função Y será dada por

$$w(y) = \frac{\bar{y}_1 - \bar{y}_2}{n_1+n_2-p-1} = \frac{0,2035}{97} = 0,0021.$$

Admitindo a normalidade de Y podemos estimar as probabilidades α e β (são iguais) de erro de classificação.



$$Y_1 \sim N(0,2930;\ 0,0021)$$

$$Y_2 \sim N(0,0895;\ 0,0021)$$

$$\alpha = P(Y_2 > \frac{\bar{y}_1+\bar{y}_2}{2}) = P(Y_2 > 0,1913) = 0,0132.$$

É imediato que se tivéssemos usado o comprimento e a largura da pétala em conjunto com estas medidas, teríamos uma probabilidade de erro muito menor visto que as características das pétalas por si só já fornecem uma boa discriminação.

* * *

Passemos agora ao quadro de Análise de Variância para testar a significância da discriminação.

FONTE	G.L.	S.Q.	Q.M.	F
ENTRE AMOSTRAS	2	1,035	0,5175	246,43
DENTRO DE AMOSTRAS	97	0,2035	0,0021	
TOTAL	99	1,2385	--	---

O valor de $F_{2,97}$ (1%) é 4,90 o que mostra que a discriminação é altamente significante.

2.8 - INTERPRETAÇÃO GEOMÉTRICA

Para melhor visualizar a interpretação geométrica da função discriminante proposta por Fisher, consideremos o caso em que o número de características associadas a cada

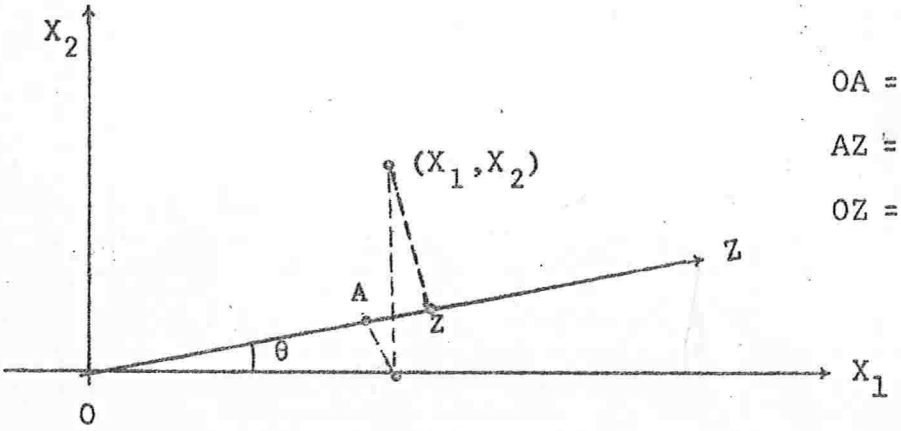
indivíduo é $p = 2$. Para um valor maior de p a noção geométrica que daremos pode ser facilmente extendida porém difícil de ser visualizada.

Em um referencial ortogonal (X_1, X_2) marquemos os pontos correspondentes aos n_1+n_2 elementos das duas amostras, utilizando sinais distintos para cada uma das populações.

Consideremos a função linear

$$Z = X_1 \cos\theta + X_2 \operatorname{sen}\theta$$

que representa a projeção do ponto (X_1, X_2) sobre um eixo OZ que forma um ângulo θ com o eixo OX_1. Mostremos que a função $Z = X_1 \cos\theta' + X_2 \operatorname{sen}\theta'$ para a qual a superposição das projeções dos pontos das duas populações é mínima, não é nada mais do que a função discriminante de Fisher.



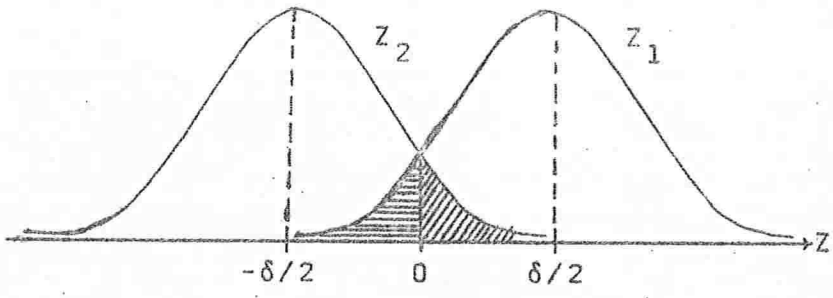
$OA = X_1 \cos\theta$

$AZ = X_2 \operatorname{sen}\theta$

$OZ = X_1 \cos\theta + X_2 \operatorname{sen}\theta$

Seja δ a distância entre as projeções das médias de π_1 e π_2; seja σ o desvio padrão da distribuição projetada. Se admitirmos π_1 e π_2 normais é evidente que $Z = X_1 \cos\theta + X_2 \operatorname{sen}\theta$ será normal. Podemos então esquematizar as distribuições das

projeções da seguinte forma:



Devido a simetria teremos $P(Z_1 < 0) = P(Z_2 > 0)$, portanto a superposição pode ser entendida como $2P(Z_1 < 0)$

Superposição Mínima $\Longleftrightarrow$ $P(Z_1 < 0)$ mínima $\Longleftrightarrow$ $(\frac{\delta}{\sigma})$ máximo.

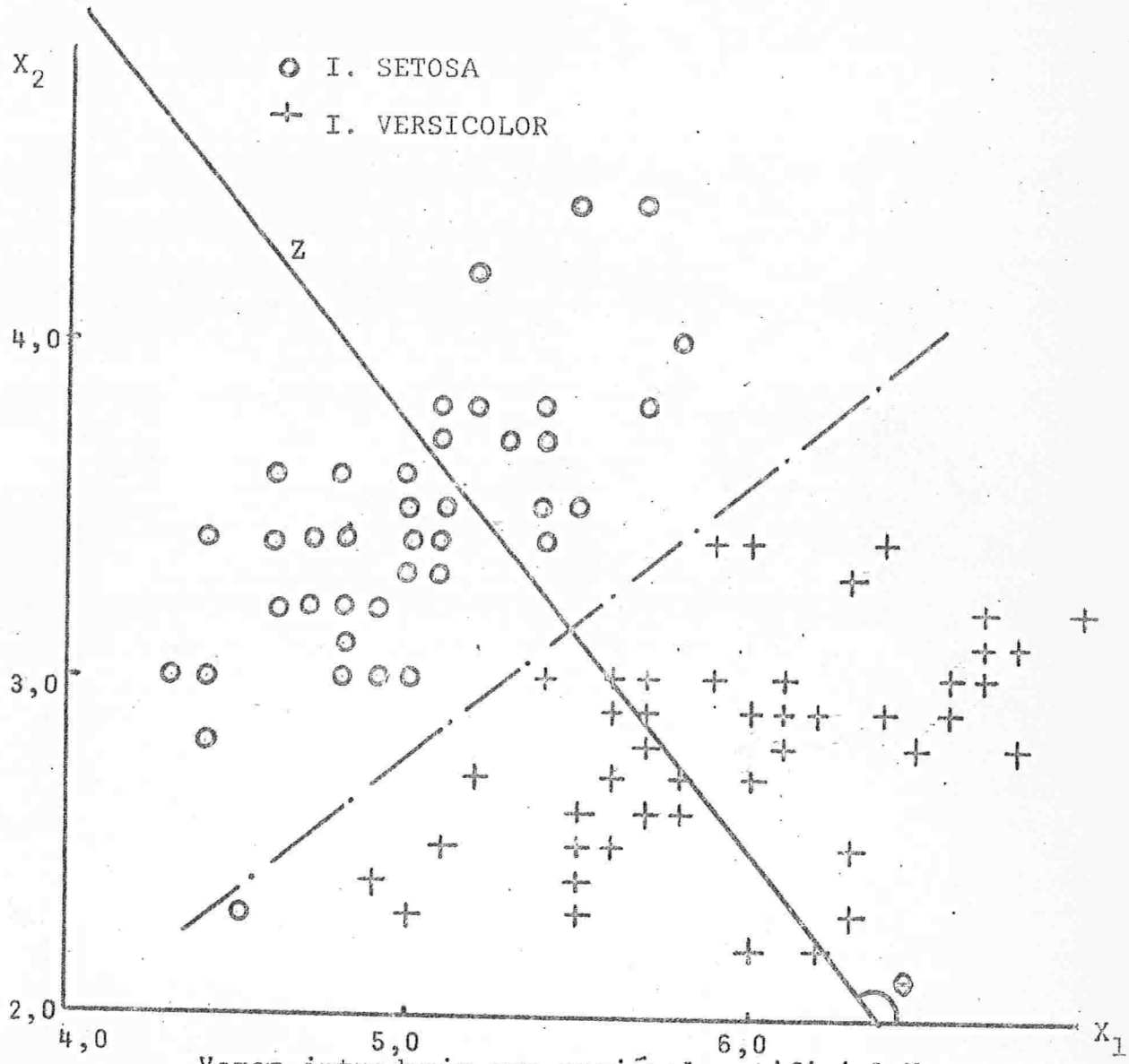
Ora (δ/σ) máximo não é nada mais do que o critério de melhor discriminação proposto por Fisher para obtenção da função discriminante. Se a função linear $Z = X_1 \cos \theta' + X_2 \operatorname{sen} \theta'$ é aquela para aqual corresponde (δ/σ) máximo esta função é a função discriminante de Fisher.

Na figura da página seguinte, temos a ilustração do exposto para os dados do exemplo do item (2.7).

2.9 - UM PARALELO COM REGRESSÃO MULTILINEAR

Um interessante paralelo é apresentado por Fisher nos seus artigos de 1936 e 1938 (referências 3 e 4). É a verificação da analogia entre os coeficientes da função discriminante e os coeficientes de regressão multilinear. Vista esta analogia podemos considerar a função discriminante

como a solução de um problema de regressão.



Vamos introduzir uma variável artificial Y que nos permita distinguir os elementos de π_1 e π_2.

— a cada um dos n_1 elementos da amostra de π_1 associaremos o valor $y = \frac{n_2}{n_1 + n_2}$

— a cada um dos n_2 elementos da amostra de π_2 associaremos o valor $y = \frac{-n_1}{n_1 + n_2}$.

É imediato que

$$\sum y = 0 \quad e \quad \sum y^2 = \frac{n_1 \, n_2}{n_1+n_2} = \lambda^2$$

onde a somatória é extendida aos n_1+n_2 elementos amostrais.

A partir desses valores de Y, procuremos estabelecer a regressão de Y sobre $X_1,\ldots,X_p$

$$y = \sum_{i=1}^{p} b^i(x_i-\bar{x}_i)$$

onde b^i representa o coeficiente da regressão associado à variável X_i,

Sabemos da regressão multilinear que os coeficientes b^i podem ser obtidos do sistema de equações

$$\sum_{j=1}^{p} s_{ij}b^j = s_{iy} \qquad (2.9.1)$$

onde

$$s_{ij} = \sum(x_i-\bar{x}_i)(x_j-\bar{x}_j)$$

$$s_{iy} = \sum(x_i-\bar{x}_i)(y-\bar{y}).$$

Como

$$d_i = \bar{x}_{1i} - \bar{x}_{2i}$$

a equação (2.9.1) pode ser reescrita assim

$$\sum_{j=1}^{p} s_{ij}b^j = \lambda^2 d_i \qquad (2.9.2)$$

Denotando por S_{ij} a soma de produtos

$$S_{ij} = \sum_1 (x_{1ik}-\bar{x}_{1i})(x_{1jk}-\bar{x}_{1j}) + \sum_2 (x_{2ik}-\bar{x}_{2i})(x_{2jk}-\bar{x}_{2j})$$

teremos

$$s_{ij} = S_{ij} + \lambda^2 d_i d_j$$

Substituindo esta última expressão em (2.9.2), teremos

$$\sum_{j=1}^{p} (S_{ij}+\lambda^2 d_i d_j) b^j = \lambda^2 d_i \Longrightarrow \sum_{j=1}^{p} S_{ij} b^j = \lambda^2 d_i (1-\sum_{j=1}^{p} b^j d_j)$$

Ora, a menos da constante $\lambda^2(1-\sum_{j=1}^{p} b^j d_j)$ teremos o mesmo sistema de equações utilizado na determinação dos coeficientes da função discriminante linear, a saber:

$$\sum_{j=1}^{p} S_{ij} b^j = d_i .$$

Comentamos anteriormente que quando o objetivo é simplesmente a obtenção da melhor discriminação, constantes que multipliquem todos os coeficientes b podem ser eliminadas ou acrescidas convenientemente. Assim, podemos supor a função discriminante como uma solução de um problema de regressão.

Fisher (referências 3.4) monta o quadro de análise de regressão e mostra que a significância da regressão é equivalente, no caso, à significância da discriminação. Assim, se a regressão for altamente significativa, também o será a discriminação.

- □ -

CAPÍTULO 3

DISCRIMINAÇÃO ENTRE DUAS POPULAÇÕES

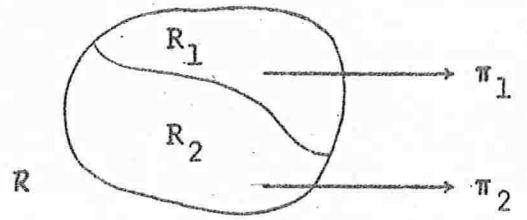
OUTROS CRITÉRIOS

3.1 - INTRODUÇÃO

Consideremos duas populações π_1 e π_2. A cada elemento dessas populações associaremos um vetor $\underline{x}' = (x_1, \ldots, x_p)$ de valores do vetor aleatório $\underline{X}' = (X_1, \ldots, X_p)$. Ao conjunto de todos os vetores $\underline{x}$ denotaremos por R sendo este um subconjunto do espaço p-dimensional $\mathbb{R}^p$. Suporemos conhecidas as duas populações π_1 e π_2, isto é as distribuições conjuntas de $X_1, \ldots, X_p$ nessas duas populações.

Seja E um elemento escolhido ao acaso dentre todos os elementos de π_1 e π_2. Nosso problema é o estabelecimento de uma regra discriminatória que nos permita da melhor forma possível classificar esse elemento E em uma das duas populações: - a sua população de origem. Para tal, particionaremos o espaço amostral R em R_1 e R_2 de forma que se o vetor $\underline{x}$ associado ao elemento E pertencer a R_1, este elemento será classificado em π_1 e caso contrário, isto é, se $\underline{x}$ per-

tencer a R_2, o elemento será classificado em π_2. Devemos portanto procurar a "melhor" partição $R = (R_1, R_2)$ no sentido de minimizar os efeitos de um erro de classificação.



Adotada uma partição R como regra discriminatória, dois tipos de erro podem ocorrer: o elemento E ser classificado em π_1 quando na realidade pertence a π_2 ou ser classificado em π_2 quando na realidade pertence a π_1. Denotaremos por

$$P(i|j,R) \quad i=1,2, \quad j=1,2$$

a probabilidade de classificar em π_i um elemento que pertence a π_j, quando adotada a regra discriminatória R.

Em muitos casos esses erros não ocorrem com a mesma frequência e além do mais as consequências que eles acarretam não são equivalentes. Para considerar estes fatores na obtenção da regra R devemos introduzir dois novos elementos:

- a probabilidade p_i de que o elemento E seja originário da população π_i (probabilidade *a priori*) sendo $p_1+p_2 = 1$.
- a função custo $C(i|j)$ que pode ser interpretada simplesmente pelo próprio nome, isto é, o "custo" incorrido ao clas

sificar um elemento de π_j em π_i. Para todos os efeitos consideraremos $C(i|i) = 0$, $i=1,2$ e $C(i|j) > 0$, $i \neq j$.

Uma maior clareza pode ser dada à introdução destes elementos com o auxílio do seguinte exemplo: — uma pessoa deve ser classificada como cardíaca (π_1) ou não cardíaca (π_2). O diagnóstico deverá ser dado por um processo discriminatório baseado em p medidas biológicas. É de se esperar que o número de cardíacos seja inferior ao de pessoas normais — estamos admitindo $p_1 < p_2$ — e se isto não for levado em conta o erro mais frequente será o de classificar uma pessoa normal como cardíaca. As consequências de uma má classificação neste caso podem ter, em um sentido, efeitos desastrosos: ao classificar um cardíaco como pessoa normal, evitando assim um tratamento adequado, pagaremos um preço bem mais alto (a vida do paciente) — maior custo — que em caso contrário quando um não cardíaco será submetido por exemplo a um intenso repouso, por ter sido classificado como cardíaco. É importante portanto que o estatístico considere os custos no processo discriminatório e como veremos mais adiante, não é necessário conhecer-se exatamente tais custos, bastando conhecer a razão entre eles.

Nos ítens seguintes consideraremos o estabelecimento da partição R em dois casos:

— as probabilidades (p_i) a priori são conhecidas

— as probabilidades a priori não são conhecidas.

3.2 - DETERMINAÇÃO DE R QUANDO AS PROBABILIDADES A PRIORI SÃO CONHECIDAS

Se conhecermos as probabilidades a priori p_i, o risco, ou seja, a perda esperada no processo discriminatório quando adotamos a partição $R = (R_1, R_2)$ será dado por

$$r(R) = C(2|1)p_1 \; P(2|1,R) + C(1|2)p_2 \; P(1|2,R)$$

Se $f_i(\underline{x}) = f_i(x_1, \ldots, x_p)$ representar a densidade conjunta de $X_1, \ldots, X_p$ na população π_i, teremos

$$r(R) = C(2|1)p_1 \int_{R_2} f_1(\underline{x})d\underline{x} + C(1|2)p_2 \int_{R_1} f_2(\underline{x})d\underline{x}$$

onde $d(\underline{x}) = dx_1 \; dx_2 \; \ldots \; dx_p$.

Entenderemos por melhor partição R aquela (ou aquelas) que minimiza (m) o risco r(R). Isto significa que R será uma solução Bayesiana para o problema.

$$r(R) = \int_{R_2} C(2|1)p_1 f_1(\underline{x})d\underline{x} + \int_{R_1} C(1|2)p_2 f_2(\underline{x})d\underline{x}$$

$$r(R) = \int_{R} C(2|1)p_1 f_1(\underline{x})d\underline{x} - \int_{R_1} C(2|1)p_1 f_1(\underline{x})d\underline{x} + \int_{R_1} C(1|2)p_2 f_2(\underline{x})d\underline{x}$$

$$r(R) = C(2|1)p_1 + \int_{R_1} \{C(1|2)p_2 f_2(\underline{x}) - C(2|1)p_1 f_1(\underline{x})\}d\underline{x} \qquad (3.2.1)$$

Como a primeira parcela do segundo membro é constante, uma partição R que minimize r(R) deve ser tal que R_1

inclua todos os vetores $\underline{x}$ para os quais o integrando da segunda parcela seja negativo e exclua todos os $\underline{x}$ para os quais o integrando seja positivo. R_1 poderá incluir ou não vetores $\underline{x}$ que anulem o integrando. Assim, uma partição R que minimiza o risco será dada por:

$$\begin{aligned} R_1&: C(1|2)p_2f_2(\underline{x}) < C(2|1)p_1f_1(\underline{x}) \\ R_2&: C(1|2)p_2f_2(\underline{x}) \geqslant C(2|1)p_1f_1(\underline{x}) \end{aligned} \qquad (3.2.2)$$

Dissemos no parágrafo anterior que R é "uma" partição que minimiza r(R). Sim, pois R não é única. Se considerarmos uma partição $R^* = (R_1^*, R_2^*)$ em que R_1^* inclua alguns ou todos os vetores $\underline{x}$ que anulam o integrando de (3.2.1), R^* será ótima pois $r(R^*) = r(R)$ e portanto mínimo. Podemos no entanto dizer que a regra discriminatória será única a menos dos vetores $\underline{x}$ para os quais $C(1|2)p_2f_2(\underline{x}) = C(2|1)p_1f_1(\underline{x})$.

Note-se que o processo discriminatório (R) descrito acima equivale a calcular para dado elemento $\underline{x}$ a razão de verossimilhança $f_1(\underline{x})/f_2(\underline{x})$ e classificar esse elemento em π_1 se essa razão for superior à constante

$$C(1|2)p_2/C(2|1)p_1$$

ou caso contrário classificá-lo em π_2.

3.3 - DETERMINAÇÃO DE R QUANDO AS PROBABILIDADES A PRIORI SÃO DESCONHECIDAS

Ao classificar um elemento de π_i, a perda esperada

será dada por:

$$r(i,R) = C(j|i)P(j|i,R) \quad i \neq j,\ i=1,2;\ j=1,2$$

quando for adotada a regra discriminatória R.

No ítem anterior minimizamos o risco do processo, isto é, determinamos R de forma tal que r(R) fosse mínimo. No entanto, devido ao desconhecimento das p_i não podemos ir direto a esse tipo de otimização. Nosso conceito de melhor regra discriminatória sofrerá uma mudança.

Consideraremos R ótima se for tal que minimiza o maior dos riscos r(i,R). Diremos então que R é uma regra minimax. A adoção deste conceito de ótimo é visivelmente apoiada em um ponto de vista conservativo, isto é, o processo discriminatório mais conveniente é aquele em que a pior das consequências (maior risco) é mínima, um critério puramente "defensivo".

Antes de passar à determinação de R vejamos alguns conceitos que serão úteis para isso.

i) a regra R será "pelo menos tão boa" quanto a regra S se

$$r(i,R) \leq r(i,S),\ \forall i$$

ii) a regra R será "melhor" que S, se para algum i tivermos

$$r(i,R) < r(i,S)$$

iii) Uma regra R é admissível se não houver regra S melhor que R

iv) uma classe de regras é "completa" se para toda regra não pertencente à classe corresponde uma dentro da classe que é melhor

v) uma classe é "completa mínima" se nenhum subconjunto próprio dessa classe for classe completa.

Voltemos à determinação de R. É claro que vamos restringir a procura no conjunto das regras admissíveis. Pode-se demonstrar (Apêndice B) que sob certas condições o conjunto de regras Bayesianas coincide com o conjunto das regras admissíveis, sendo portanto uma classe completa mínima. Procuremos portanto a regra Bayesiana que minimiza o maior dos riscos $r(i,R)$.

Como vimos anteriormente, a regra $R = (R_1, R_2)$ é dita Bayesiana se minimiza o valor de $r(R)$ para um dado p_1 $(p_2 = 1 - p_1)$

$$r(R) = C(2|1)p_1 \; P(2|1,R) + C(1|2)p_2 \; P(1|2,R)$$

Vimos também que tais regras são assim definidas

$$R_1: \; C(2|1)p_1 f_1(\underline{x}) \geq C(1|2)p_2 f_2(\underline{x})$$

$$R_2: \; C(1|2)p_2 f_2(\underline{x}) \geq C(2|1)p_1 f_1(\underline{x})$$

O que faremos para obter a solução minimax é o seguinte: introduziremos artificialmente valores de p_1 (e consequentemente de $p_2 = 1 - p_1$). Variando os valores de p_1 obteremos as diferentes regras Bayesianas $R = (R_1, R_2)$. Faremos isso até obter um valor de p_1 que denotaremos por p_1^* tal que

para a regra R^* a ele associada tenhamos $r(1,R^*) = r(2,R^*)$. Esta evidentemente será a solução minimax por nós procurada.

É importante notar que sempre existe p^* que satisfaça a essa condição. De fato, quando $p_1 = 0$ teremos

$$P(1|2,R) = 0 \text{ e } P(2|1,R) = 1.$$

À medida que p_1 vai continuamente crescendo para 1, $P(1|2,R)$ irá crescendo continuamente para 1 e $P(2|1,R)$ decrescendo para 0. Quando $p_1 = 1$ teremos $P(1|2,R) = 1$ e $P(2|1,R) = 0$. Em outras palavras quando p_1 varia continuamente de 0 a 1, o risco $r(1,R)$ varia de $C(2|1)$ a 0, o risco $r(2,R)$ de 0 a $C(1|2)$ e a variação dos riscos também será contínua. Existe portanto um valor p^* para o qual os riscos se "cruzam" isto é $r(1,R^*) = r(2,R^*)$.

Devemos mostrar que esta solução é minimax. De fato, se houvesse uma outra regra S tal que

$$\text{máx}\{r(1,S), r(2,S)\} < r(1,R^*) = r(2,R^*)$$

isto negaria o fato de que a regra Bayesiana R^* é admissível.

A regra R^* assim obtida é denominada "regra Bayesiana a riscos constantes". O processo minimax nem sempre é utilizado por ser muito trabalhoso (especialmente para o caso de mais de duas populações) sendo substituído pela adoção da regra Bayesiana R correspondente à igualdade das probabilidades a priori ($p_1 = p_2 = 0,5$). Aliás, este útimo processo ($p_1 = p_2 = 0,5$) é sugerido por Rao (referência 9) para o

caso em que as probabilidades a priori são desconhecidas.

3.4 - APLICAÇÃO PARA O CASO DE POPULAÇÕES NORMAIS

Neste ítem iremos aplicar os resultados obtidos nos ítens 3.2 e 3.3 para o caso em que as duas populações consideradas π_1 e π_2, são normais multivariadas com mesma matriz de covariância. Teremos:

$$\pi_1 \ldots (X_1, \ldots, X_p) \sim N_p(\underline{\mu}_1, \Sigma)$$

$$\pi_2 \ldots (X_1, \ldots, X_p) \sim N_p(\underline{\mu}_2, \Sigma)$$

onde

$\underline{\mu}_i' = (\mu_{i1}, \ldots, \mu_{ip})$ $i=1,2$ vetor das médias

Σ: matriz de covariância (não singular).

A função densidade de probabilidade de $(X_1,\ldots,X_p)$ na população π_i será dada por:

$$f_i(\underline{x}) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\{-\frac{1}{2}(\underline{x}-\underline{\mu}_i)'\Sigma^{-1}(\underline{x}-\underline{\mu}_i)\}$$

onde $|\Sigma|$ é o determinante de Σ e $f_i(\underline{x}) = f_i(x_1,\ldots,x_p)$.

Como vimos nos ítens anteriores, as regiões R_1 e R_2 ficam definidas por

$$R_1: \frac{f_1(\underline{x})}{f_2(\underline{x})} \geq k \qquad R_2: \frac{f_1(\underline{x})}{f_2(\underline{x})} \leq k$$

sendo k uma constante convenientemente escolhida. Analisaremos qual deve ser essa constante k nas diferentes situações. Antes porém, considerando que quando $f_1(\underline{x})/f_2(\underline{x})$ cresce seu logaritmo também cresce, podemos escrever R_1 e R_2 da seguinte forma:

$$R_1: \log \frac{f_1(\underline{x})}{f_2(\underline{x})} \geq \log k \qquad R_2: \log \frac{f_1(\underline{x})}{f_2(\underline{x})} \leq \log k$$

onde "log" representa o logaritmo neperiano.

A razão entre densidades resulta

$$\frac{f_1(\underline{x})}{f_2(\underline{x})} = \frac{\exp\{-\frac{1}{2}(\underline{x}-\underline{\mu}_1)'\Sigma^{-1}(\underline{x}-\underline{\mu}_1)\}}{\exp\{-\frac{1}{2}(\underline{x}-\mu_2)'\Sigma^{-1}(\underline{x}-\underline{\mu}_2)\}}$$

ou

$$\frac{f_1(\underline{x})}{f_2(\underline{x})} = \exp\{-\frac{1}{2}(\underline{x}-\underline{\mu}_1)'\Sigma^{-1}(\underline{x}-\underline{\mu}_1)+\frac{1}{2}(\underline{x}-\underline{\mu}_2)'\Sigma^{-1}(\underline{x}-\underline{\mu}_2)\}.$$

Assim, as regiões R_1 e R_2 ficam definidas por

$$R_1: -\frac{1}{2}\{(\underline{x}-\underline{\mu}_1)'\Sigma^{-1}(\underline{x}-\underline{\mu}_1)-(\underline{x}-\underline{\mu}_2)'\Sigma^{-1}(\underline{x}-\underline{\mu}_2)\} \geq \log k$$

$$R_2: -\frac{1}{2}\{(\underline{x}-\underline{\mu}_1)'\Sigma^{-1}(\underline{x}-\underline{\mu}_1)-(\underline{x}-\underline{\mu}_2)'\Sigma^{-1}(\underline{x}-\underline{\mu}_2)\} \leq \log k$$

Desenvolvendo o primeiro membro das desigualdades acima e reagrupando convenientemente teremos

$$R_1: \underline{x}'\Sigma^{-1}(\underline{\mu}_1-\underline{\mu}_2)-\frac{1}{2}(\underline{\mu}_1+\underline{\mu}_2)'\Sigma^{-1}(\underline{\mu}_1-\underline{\mu}_2) \geq \log k$$

$$R_2: \underline{x}'\Sigma^{-1}(\underline{\mu}_1-\underline{\mu}_2)-\frac{1}{2}(\underline{\mu}_1+\underline{\mu}_2)'\Sigma^{-1}(\underline{\mu}_1-\underline{\mu}_2) \leq \log k.$$

A primeira parcela do primeiro termo das desigualdades é denominada "função discriminante linear" por ser uma função linear dos componentes de $\underline{x}' = (x_1, \ldots, x_p)$ podendo ser vista como o modelo populacional da função discriminante proposta por Fisher. A segunda parcela do primeiro termo não depende do vetor de observações, é uma constante que depende apenas das distribuições das duas populações. O segundo membro é uma constante cujo valor depende do conhecimento ou não das probabilidades a priori. Examinemos os dois casos:

a) As probabilidades a priori são conhecidas

Neste caso temos de acordo com o ítem 3.2

$$k = \frac{C(1|2)p_2}{C(2|1)p_1}$$

No caso particular em que $p_1 = p_2$ e $C(1|2) = C(2|1)$ teremos $k = 1$ e $\log k = 0$, ficando R_1 e R_2 assim definidos:

$$R_1: \quad \underline{x}'\Sigma^{-1}(\underline{\mu}_1 - \underline{\mu}_2) \geq \frac{1}{2}(\underline{\mu}_1 + \underline{\mu}_2)'\Sigma^{-1}(\underline{\mu}_1 - \underline{\mu}_2)$$

$$R_2: \quad \underline{x}'\Sigma^{-1}(\underline{\mu}_1 - \underline{\mu}_2) \leq \frac{1}{2}(\underline{\mu}_1 + \underline{\mu}_2)'\Sigma^{-1}(\underline{\mu}_1 - \underline{\mu}_2)$$

b) As probabilidades a priori não são conhecidas

Se não conhecermos as probabilidades a priori, devemos conforme o ítem 3.3 recorrer ao método minimax o que corresponde à determinação do valor de k para o qual haja igualdade dos riscos $r(i,R)$. No cálculo dos riscos devemos conhecer as probabilidades de erro $P(i|j,R)$ e para isto é ne

cessário conhecer a função densidade de probabilidade da variável

$$U = \underline{X}'\Sigma^{-1}(\underline{\mu}_1-\underline{\mu}_2) - \frac{1}{2}(\underline{\mu}_1+\underline{\mu}_2)'\Sigma^{-1}(\underline{\mu}_1-\underline{\mu}_2) .$$

$P(U \leq \log k|\pi_1)$ representa $P(2|1,R)$ ou seja a probabilidade de classificar erroneamente em π_2 um elemento que pertence a π_1. Reciprocamente, $P(U \geq \log k|\pi_2)$ representa $P(1|2,R)$.

Se $\underline{X}' = (X_1,\ldots,X_p)$ tem distribuição normal multivariada, a variável U, por ser uma combinação linear de $X_1,\ldots,X_p$, terá distribuição normal unidimensional.

Se $\underline{X} \sim N_p(\underline{\mu}_1,\Sigma)$ então a variável U terá os parâmetros

$$E_1(U) = \underline{\mu}_1'\Sigma^{-1}(\underline{\mu}_1-\underline{\mu}_2) - \frac{1}{2}(\underline{\mu}_1+\underline{\mu}_2)'\Sigma^{-1}(\underline{\mu}_1-\underline{\mu}_2) =$$

$$= \frac{1}{2}(\underline{\mu}_1-\underline{\mu}_2)'\Sigma^{-1}(\underline{\mu}_1-\underline{\mu}_2).$$

$$Var_1(U) = E_1\{U-E_1(U)\}^2 = E_1\{(\underline{X}-\underline{\mu}_1)'\Sigma^{-1}(\underline{\mu}_1-\underline{\mu}_2)\}^2 =$$

$$= (\underline{\mu}_1-\underline{\mu}_2)'\Sigma^{-1}\Sigma\Sigma^{-1}(\underline{\mu}_1-\underline{\mu}_2) = (\underline{\mu}_1-\underline{\mu}_2)'\Sigma^{-1}(\underline{\mu}_1-\underline{\mu}_2)$$

Se $\underline{X} \sim N_p(\underline{\mu}_2,\Sigma)$ então a variável U terá os parâmetros

$$E_2(U) = -\frac{1}{2}(\underline{\mu}_1-\underline{\mu}_2)'\Sigma^{-1}(\underline{\mu}_1-\underline{\mu}_2)$$

$$Var_2(U) = (\underline{\mu}_1-\underline{\mu}_2)'\Sigma^{-1}(\underline{\mu}_1-\underline{\mu}_2).$$

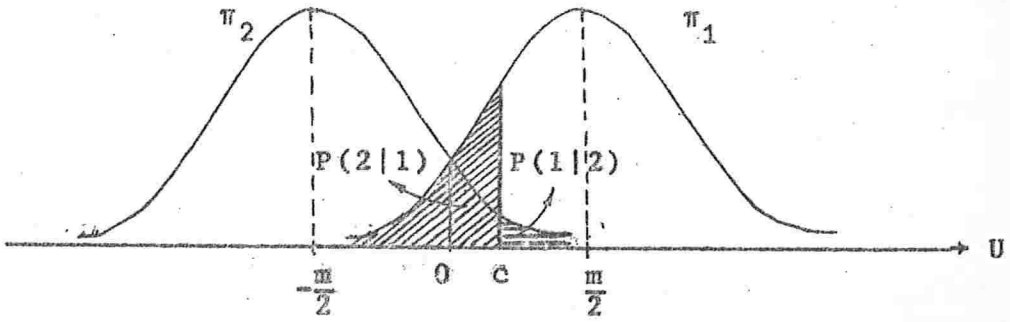
Note-se que $E_1(U) - E_2(U) = (\underline{\mu}_1-\underline{\mu}_2)'\Sigma^{-1}(\underline{\mu}_1-\underline{\mu}_2)$ que nada mais é do que a distância generalizada de Mahalanobis baseada nas p variáveis $X_1,\ldots,X_p$ (Vide Apêndice A).

Se denotarmos $(\underline{\mu}_1-\underline{\mu}_2)'\Sigma^{-1}(\underline{\mu}_1-\underline{\mu}_2) = m$ (de Mahalanobis) teremos

$$\pi_1 \ldots \; U \sim N(\tfrac{m}{2},m)$$

$$\pi_2 \ldots \; U \sim N(-\tfrac{m}{2},m)$$

o que fornece o seguinte esquema



Se denotarmos logk = c teremos as probabilidades de erro dadas por:

$$P(2|1) = \int_{-\infty}^{c} \frac{1}{\sqrt{2\pi m}} e^{-\frac{1}{2}(u-\frac{m}{2})^2/m} du = \int_{-\infty}^{(c-\frac{m}{2})/\sqrt{m}} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2} dy$$

e

$$P(1|2) = \int_{c}^{\infty} \frac{1}{\sqrt{2\pi m}} e^{-\frac{1}{2}(u+\frac{m}{2})^2/m} du = \int_{(c+\frac{m}{2})/\sqrt{m}}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2} dy.$$

O valor c que determina a regra R minimax será obtido resolvendo-se a equação:

$$C(1|2)P(1|2) = C(2|1)P(2|1)$$

ou seja igualando os riscos. As regiões R_1 e R_2 serão dadas por:

$$R_1 : u \geq c$$

$$R_2 : u \leq c$$

No caso particular $C(1|2) = C(2|1)$ teremos $c = 0$ como solução e portanto

$$R_1 : \underline{x}'\Sigma^{-1}(\underline{\mu}_1 - \underline{\mu}_2) \geq \frac{1}{2}(\underline{\mu}_1 + \underline{\mu}_2)'\Sigma^{-1}(\underline{\mu}_1 - \underline{\mu}_2)$$

coincidindo com o resultado obtido anteriormente quando, além dos custos, as probabilidades a priori (conhecidas) também eram iguais.

3.4.1 - Caso em que os parâmetros $\underline{\mu}$ e Σ são desconhecidos

Como vimos anteriormente a regra discriminatória

$$R = (R_1, R_2)$$

é definida por:

$$R_1 : \underline{x}'\Sigma^{-1}(\underline{\mu}_1 - \underline{\mu}_2) - \frac{1}{2}(\underline{\mu}_1 + \underline{\mu}_2)'\Sigma^{-1}(\underline{\mu}_1 - \underline{\mu}_2) \geq \log k$$

$$R_2 : \underline{x}'\Sigma^{-1}(\underline{\mu}_1 - \underline{\mu}_2) - \frac{1}{2}(\underline{\mu}_1 + \underline{\mu}_2)'\Sigma^{-1}(\underline{\mu}_1 - \underline{\mu}_2) \leq \log k.$$

Em termos de aplicação da técnica uma das limitações que surge é o desconhecimento dos parâmetros populacionais $\underline{\mu}_1$, $\underline{\mu}_2$, Σ. A saída intuitivamente lógica é a substituição desses parâmetros por seus melhores estimadores. É claro que isto introduzirá novos erros de discriminação, tanto maiores quanto mais pobre o material sobre o qual se basear a estimação.

Sabemos que as melhores estimativas de $\underline{\mu}_i$ e Σ são

respectivamente $\bar{\underline{x}}_i$ e W.

O vetor $\bar{\underline{x}}_i$ é definido por:

$$\bar{\underline{x}}_i = \frac{1}{n_i} \sum_{k=1}^{n_i} \underline{x}_{ik}$$

sendo n_i o tamanho da amostra e $\underline{x}_{ik}$ o vetor que representa o k-ésimo elemento da amostra da população π_i.

A matriz W é definida por

$$(n_1+n_2-2)W = \sum_{k=1}^{n_1} (\underline{x}_{1k}-\bar{\underline{x}}_1)(\underline{x}_{1k}-\bar{\underline{x}}_1)' + \sum_{k=1}^{n_2} (\underline{x}_{2k}-\bar{\underline{x}}_2)(\underline{x}_{2k}-\bar{\underline{x}}_2)'$$

Substituindo $\underline{\mu}_i$ e Σ por suas estimativas $\bar{\underline{x}}_i$ e W teremos as regiões R_1 e R_2 assim definidas

$$R_1: \underline{x}'W^{-1}(\bar{\underline{x}}_1-\bar{\underline{x}}_2) - \frac{1}{2}(\bar{\underline{x}}_1+\bar{\underline{x}}_2)'W^{-1}(\bar{\underline{x}}_1-\bar{\underline{x}}_2) \geq \log k$$

$$R_2: \underline{x}'W^{-1}(\bar{\underline{x}}_1-\bar{\underline{x}}_2) - \frac{1}{2}(\bar{\underline{x}}_1+\bar{\underline{x}}_2)'W^{-1}(\bar{\underline{x}}_1-\bar{\underline{x}}_2) \leq \log k$$

Note-se que a primeira parcela do primeiro termo nada mais é do que a função discriminante linear proposta por Fisher (Capítulo 2).

Para avaliarmos o risco quando da adoção dos estimadores devemos conhecer as probabilidades de cometer um erro de classificação o que só se torna possível com o conhecimento da distribuição de

$$V = \underline{X}'W^{-1}(\bar{\underline{X}}_1-\bar{\underline{X}}_2) - \frac{1}{2}(\bar{\underline{X}}_1+\bar{\underline{X}}_2)'W^{-1}(\bar{\underline{X}}_1-\bar{\underline{X}}_2)$$

A distribuição de V é extremamente complicada, dependendo do tamanho das amostras e do parâmetro:

$$m = (\underline{\mu}_1 - \underline{\mu}_2)' \Sigma^{-1} (\underline{\mu}_1 - \underline{\mu}_2).$$

Esta distribuição foi inicialmente estudada por Wald (1944) porém os resultados por ele obtidos eram bastante complicados em termos de aplicação. Mais tarde, Harter (1950) tentou a simplificação dos resultados de Wald, com vistas a facilitar a aplicação. Anderson (1951) e Sitgreaves (1952) também estudaram a distribuição de V.

No entanto, como $\bar{\underline{X}}_i$ converge em probabilidade para $\underline{\mu}_i$ e W para Σ, a distribuição limite de V será a distribuição de U. Isto é garantido pelo teorema de Slutsky (cf. H. Cramer - Mathematical Methods of Statistics) sobre a convergência em probabilidade de funções racionais de variáveis aleatórias que convergem em probabilidade a determinadas constantes.

O exposto no parágrafo acima significa em termos de aplicação que para amostras suficientemente grandes a utilização das estimativas de $\underline{\mu}_i$ e Σ não introduz grandes erros no processo de classificação.

3.5 - RELACIONAMENTO DA FUNÇÃO DISCRIMINANTE DE FISHER COM A SOLUÇÃO BAYESIANA

O objetivo deste ítem é relacionar a função discriminante de Fisher com as soluções obtidas no presente capítulo e assim mostrar que no caso particular de populações normais com mesma matriz de covariância os resultados coin-

cidem, isto é, a função discriminante proposta por Fisher define a melhor partição R do espaço amostral.

Como vimos no capítulo anterior, a função discriminante obtida por Fisher é

$$y = \underline{x}'S^{-1}(\bar{\underline{x}}_1 - \bar{\underline{x}}_2)$$

onde $S^{-1}(\bar{\underline{x}}_1 - \bar{\underline{x}}_2)$ é o vetor dos coeficientes b_i da função, sendo S a matriz da soma de quadrados e produtos dentro das amostras, portanto $(n_1+n_2-2)W = S$.

Comentamos também que a multiplicação ou divisão dos coeficientes por uma constante não nula em nada alterava o efeito discriminatório da função. Assim sendo preferiremos neste ítem escrever a função discriminante da seguinte forma

$$y = \underline{x}'W^{-1}(\bar{\underline{x}}_1 - \bar{\underline{x}}_2)$$

em que os novos coeficientes são os anteriores multiplicados por n_1+n_2-2.

Examinaremos os dois casos seguintes: probabilidades a priori desconhecidas e probabilidades a priori conhecidas.

a) Probabilidades a Priori Desconhecidas

A solução minimax apresentada no ítem 3.4. (caso b) recaía na determinação de um ponto c de corte tal que se $u \geq c$ o elemento era classificado em π_1 e se $u \leq c$ em π_2.

Como se trata de um caso em que contaremos apenas com as estimativas dos parâmetros populacionais, trabalharemos com a estatística V e desta forma teremos

$$R_1: v \geq c \quad e \quad R_2: v \leq c$$

Vimos também que no caso dos custos $C(1|2)$ e $C(2|1)$ serem iguais (admitido implicitamente no trabalho de Fisher) o valor de c é 0 (zero). Assim

$$R_1: \underline{x}'W^{-1}(\bar{\underline{x}}_1-\bar{\underline{x}}_2) - \frac{1}{2}(\bar{\underline{x}}_1+\bar{\underline{x}}_2)'W^{-1}(\bar{\underline{x}}_1-\bar{\underline{x}}_2) \geq 0$$

ou

$$R_1: \underline{x}'W^{-1}(\bar{\underline{x}}_1-\bar{\underline{x}}_2) \geq \frac{1}{2}(\bar{\underline{x}}_1+\bar{\underline{x}}_2)'W^{-1}(\bar{\underline{x}}_1-\bar{\underline{x}}_2)$$

Mas

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

e portanto

$$R_1: \underline{x}'W^{-1}(\bar{\underline{x}}_1-\bar{\underline{x}}_2) \geq \frac{1}{2}(\bar{y}_1+\bar{y}_2) \quad \text{ou} \quad y \geq \frac{1}{2}(\bar{y}_1+\bar{y}_2)$$

$$R_2: \underline{x}'W^{-1}(\bar{\underline{x}}_1-\bar{\underline{x}}_2) \leq \frac{1}{2}(\bar{y}_1+\bar{y}_2) \quad \text{ou} \quad y \leq \frac{1}{2}(\bar{y}_1+\bar{y}_2)$$

exatamente como o proposto por Fisher.

b) <u>Probabilidades a Priori Conhecidas</u>

Mostraremos agora que, embora continuemos utilizando a função discriminante proposta por Fisher, se as probabilidades p_1 e p_2 forem conhecidas o melhor ponto de corte não será mais $\frac{1}{2}(\bar{y}_1+\bar{y}_2)$ como propos Fisher, mas sim

$$\frac{1}{2}(\bar{y}_1+\bar{y}_2) + \log p_2 - \log p_1$$

De fato, de acordo com o ítem 3.4.1

$$R_1 : \underline{x}'W^{-1}(\bar{\underline{x}}_1-\bar{\underline{x}}_2)-\frac{1}{2}(\bar{\underline{x}}_1+\bar{\underline{x}}_2)'W^{-1}(\bar{\underline{x}}_1-\bar{\underline{x}}_2) \geq \log k$$

onde a constante k é dada por

$$k = \frac{p_2 C(1|2)}{p_1 C(2|1)}$$

Supondo os custos iguais, teremos

$$\log k = \log p_2 - \log p_1$$

e portanto

$$R_1 : \underline{x}'W^{-1}(\bar{\underline{x}}_1-\bar{\underline{x}}_2) \geq \frac{1}{2}(\bar{y}_1+\bar{y}_2) + \log p_2 - \log p_1 .$$

Note-se que com isto deslocamos o ponto de corte em direção à população que apresentar menor probabilidade a priori, o que é bastante lógico intuitivamente. No caso particular de $p_1 = p_2$ obteremos exatamente o ponto de corte proposto por Fisher.

3.6 - A FUNÇÃO DISCRIMINANTE QUADRÁTICA

No caso em que as populações π_1 e π_2 se distribuem normalmente mas com diferentes matrizes de covariância, Σ_1 e Σ_2 respectivamente, o logaritmo da razão de verossimilhança conduz a uma forma quadrática na definição de R_1, a saber

$$R_1 : -\frac{1}{2}\{(\underline{x}-\underline{\mu}_1)'\Sigma_1^{-1}(\underline{x}-\underline{\mu}_1) + (\underline{x}-\underline{\mu}_2)'\Sigma_2^{-1}(\underline{x}-\underline{\mu}_2)\} \geq \log k'$$

Isto significa que a superfície discriminadora é uma "quádrica" no espaço p-dimensional. Smith (1947) introduziu e empregou as funções discriminantes quadráticas. Tais tipos de funções são de difícil manipulação motivo pelo qual esse campo não se tem desenvolvido satisfatoriamente.

3.7 - A REGIÃO DE DÚVIDA DE RAO

O processo discriminatório abordado até agora se resume à partição do espaço amostral em duas regiões R_1 e R_2 de forma a minimizar os riscos do processo. Os cálculos são baseados nos valores de p variáveis associadas a cada elemento.

É claro que um aumento de número p dessas variáveis levará em geral a uma melhoria na discriminação. No entanto esse aumento de variáveis, em termos de aplicação, não só dificultará os cálculos como pode representar um sensível aumento do custo operacional. Ideal seria proceder à consideração de um número maior de variáveis apenas no caso em que for grande a dúvida sobre a verdadeira população de origem do elemento a ser classificado.

Por exemplo, um médico interessado em diagnosticar a necessidade ou não de submeter um paciente a uma intervenção cirúrgica, submete-o a certos testes. Se as respostas a estes testes não lhe dão a necessária segurança no diagnóstico (a probabilidade de erro supera um máximo permissível)

ele pode submeter o paciente a outros testes visando o melhor diagnóstico possível, isto é, com a menor chance de erro.

Com o objetivo de atender casos desta espécie, em que desejamos controlar as probabilidades de erro em limites pré-estabelecidos, podemos utilizar o critério discriminante exposto por C.R.Rao (referência 9), que denominaremos "critério da região de dúvida" ou simplesmente "critério de Rao".

Basicamente, o critério de Rao consiste em particionar o espaço amostral em 3 regiões, R_1, R_2 e R_0. Se o vetor $\underline{x}$ associado ao elemento E pertencer a R_1, E será classificado em π_1; se pertencer a R_2, E será classificado em π_2; se no entanto pertencer a R_0, novas características X_{p+1}, X_{p+2}, ... deverão ser consideradas para classificar o elemento em π_1 ou π_2. A região R_0 é denominada "região de dúvida".

O critério para definição das regiões R_0, R_1, R_2 é assim dado por Rao:

$$R_1: f_1(\underline{x}) \geq Af_2(\underline{x})$$

$$R_2: f_1(\underline{x}) \leq Bf_2(\underline{x})$$

$$R_0: Bf_2(\underline{x}) < f_1(\underline{x}) < Af_2(\underline{x})$$

onde os valores de A e B são determinados em função dos limites de probabilidade de erro pré-estabelecidos.

Por exemplo, consideremos duas populações normais com mesma matriz de covariância, e sejam α_1 e α_2 os valores máximos desejados de $P(1|2,R)$ e $P(2|1,R)$ respectivamente onde

$$R = (R_1, R_2, R_0)$$

é o critério de Rao.

Lembrando que

$$U = \underline{X}'\Sigma^{-1}(\underline{\mu}_1 - \underline{\mu}_2) - \frac{1}{2}(\underline{\mu}_1 + \underline{\mu}_2)'\Sigma^{-1}(\underline{\mu}_1 - \underline{\mu}_2)$$

podemos escrever as regiões R_1, R_2 e R_0 da seguinte forma

R_1: $u \geq A'$

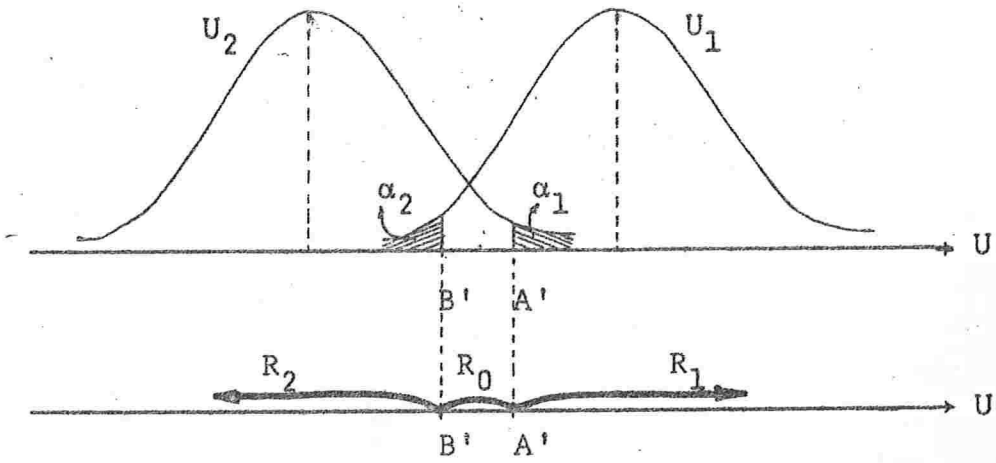
R_2: $u \leq B'$ onde $A' = \log A$ e $B' = \log B$

R_0: $B' < u < A'$

sendo A' e B' dados por

$$\int_{A'}^{\infty} u_2 du_2 = \alpha_1 \quad e \quad \int_{-\infty}^{B'} u_1 du_1 = \alpha_2$$

onde U_i é a variável U admitindo que X pertence a π_i. Esquematicamente teremos:



3.8 - OUTROS CRITÉRIOS DISCRIMINANTES

É evidente que podem existir outros critérios discriminantes além dos expostos. Por exemplo, em Anderson (referência 1) encontramos o critério da razão de verossimilhança: o problema é colocado na forma de um teste de hipóteses, a região crítica deste teste é determinada (razão de verossimilhança) e esta região é tomada por exemplo como R_1 (depende da hipótese de nulidade) e a região de aceitação como R_2.

Em Romeder (referência 12) temos o critério de Sebestyen que basicamente consiste em definir um índice, $S(E, \pi_i)$, "similaridade", tal que este índice, que é função de uma métrica previamente determinada, nos dê uma ideia da semelhança de E com os elementos de π_i. Calcula-se esse índice para todas as populações (o processo é válido para mais de 2 populações) e classifica-se E na população com a qual sua semelhança seja maior.

3.9 - TESTES RELACIONADOS COM A DISCRIMINAÇÃO ENTRE DUAS POPULAÇÕES

Apresentamos neste ítem 2 testes que nos parecem importantes, relacionados com a Análise Discriminante. Admitiremos a normalidade das populações. Utilizaremos o Apêndice A.

3.9.1 - Teste para a Diferença de Médias entre Duas Populações

Este teste é uma extensão do teste "t" de Student para duas populações normais unidimensionais. O teste se baseia no uso da estatística T^2 introduzida por Hotelling em 1931. A importância deste teste é clara na sua aplicação à Análise Discriminante: se as duas populações tiverem mesma média não tem sentido proceder à discriminação entre elas. Importante notar que este teste coincide com a Análise de Variância proposta por Fisher (referência 3).

Dadas as populações π_1 e π_2 e o vetor $\underline{X} = (X_1, \ldots, X_p)$ cuja distribuição sob a população π_i é dada por

$$\underline{X} \sim N_p(\underline{\mu}_i, \Sigma)$$

queremos testar

$$H_0\{\underline{\mu}_1 = \underline{\mu}_2 \quad \text{vs.} \quad H_a\{\underline{\mu}_1 \neq \underline{\mu}_2$$

A estatística a ser usada para o teste é (cf. Anderson, ref. 1):

$$F = \frac{n_1+n_2-p-1}{(n_1+n_2-2)p} T^2$$

que sob a hipótese de nulidade tem distribuição F de Snedecor com p e n_1+n_2-p-1 graus de liberdade.

Conforme exposto no Apêndice A, o T^2 de Hotelling pode ser relacionado com o D^2 de Mahalanobis pela expres-

são

$$T^2 = \frac{n_1 \, n_2}{n_1+n_2} D^2$$

do que resulta que a estatística F pode ser escrita

$$F = \frac{n_1 n_2 (n_1+n_2-p-1)}{(n_1+n_2)p(n_1+n_2-2)} D^2$$

O cálculo de D^2 pode ser feito diretamente pela sua definição, isto é

$$D^2 = \sum_{i=1}^{p} \sum_{j=1}^{p} w^{ij} (\bar{x}_{1i}-\bar{x}_{2i})(\bar{x}_{1j}-\bar{x}_{2j}),$$

mas se a função discriminante linear

$$Y = \underline{X}' W^{-1} (\underline{\bar{x}}_1 - \underline{\bar{x}}_2)$$

for conhecida, o cálculo de D^2 será facilitado pois

$$D^2 = \sum_{i=1}^{p} \sum_{j=1}^{p} w^{ij} (\bar{x}_{1i}-\bar{x}_{2i})(\bar{x}_{1j}-\bar{x}_{2j}) = \sum_{i=1}^{p} b_i (\bar{x}_{1i}-\bar{x}_{2i})$$

onde b_i é o coeficiente de X_i na função discriminante.

O exemplo seguinte, apresentado em Rao (referência 10), baseia-se nos dados apresentados por Fisher em seu trabalho (referência 3) sobre a discriminação entre Iris Versicolor e Iris Setosa já citado no Capítulo 2.

MÉDIAS AMOSTRAIS

CARACTERÍSTICAS	I.VERSICOLOR	I.SETOSA	DIFERENÇA
Comp. do Caule (X_1)	5,936	5,006	0,930
Larg. do Caule (X_2)	2,770	3,428	-0,658
Comp. da Pétala (X_3)	4,260	1,462	2,789
Larg. da Pétala (X_4)	1,326	0,246	1,080

Matriz de Covariância (estimativa)

$$W = \begin{pmatrix} 0,195340 & 0,092200 & 0,099626 & 0,033055 \\ & 0,121079 & 0,047175 & 0,025251 \\ & & 0,125488 & 0,039586 \\ & & & 0,025106 \end{pmatrix}$$

Com estes elementos podemos então determinar a função discriminante. Obteremos $Y = \underline{X}' W^{-1} (\bar{\underline{X}}_1 - \bar{\underline{X}}_2)$

$$Y = -3,0692X_1 - 18,0006X_2 + 21,7641X_3 + 30,7549X_4$$

e portanto $D^2 = \sum_{i=1}^{p} b_i (\bar{x}_{1i} - \bar{x}_{2i})$

$$D^2 = (-3,0692)(0,930) + \ldots + (30,7549)(1,080) = 103,2119$$

Com este valor de D^2 obteremos

$$F = \frac{50 \times 50 \times 95}{(50+50) \times 4 \times 98} \times 103,3119 = 625,3256$$

O valor crítico de F ao nível de significância de 1% será dado por $F_{4,95} = 3,53$ o que mostra que a diferença entre médias é altamente significativa.

3.9.2 - Teste da Melhoria da Discriminação

Uma questão que pode surgir na resolução de um problema de discriminação é a seguinte: estamos trabalhando com p características $X_1,\ldots,X_p$ ao tentar estabelecer uma regra discriminante. Será que a utilização de q variáveis adicionais $X_{p+1},\ldots,X_{p+q}$ não melhoraria a discriminação?

Por exemplo, ao discriminar entre dois tipos de plantas podemos estar nos baseando em p medidas relativas ao caule. Será que a utilização de q medidas adicionais relativas à pétala não irá melhorar a discriminação?

Para esse fim, ou seja testar a melhoria da discriminação devido ao incremento do número de variáveis, apresentamos o seguinte teste (cf. Rao, ref. 9): Seja D_v^2 a estimativa da distância generalizada de Mahalanobis para v variáveis consideradas. A estatística a ser utilizada é

$$U_{p,q} = \frac{(n_1+n_2-p-q-1)n_1n_2(D_{p+q}^2-D_p^2)}{q\{(n_1+n_2)(n_1+n_2-2)+n_1n_2D_p^2\}}$$

que sob a hipótese de não melhoria da discriminação tem distribuição F com q e $n_1+n_2-p-q-1$ graus de liberdade.

Se não aplicamos diretamente o teste para $D_{p+q}^2-D_p^2$ é porque a distribuição desta diferença envolve o parâmetro populacional Δ_p^2 (distância generalizada de Mahalanobis) e se este não for conhecido não podemos aplicar o teste exato. No entanto, para amostras suficientemente grandes e Δ_p^2 suficiente

mente pequeno, teoricamente nulo, podemos considerar a distribuição de $D^2_{p+q} - D^2_p$ como independente de Δ^2_p. Neste caso a estatística

$$\frac{(n_1+n_2-p-1)(n_1+n_2-p-q-1)}{q(n_1+n_2-1)} W_{p,q}$$

onde

$$W_{p,q} = \frac{n_1 n_2}{(n_1+n_2)(n_1+n_2-2)}(D^2_{p+q}-D^2_p)$$

pode ser aproximada por uma F de Snedecor com q e $n_1 + n_2 - p - q - 1$ graus de liberdade (em geral o uso da estatística acima superestima a significância).

Um exemplo da aplicação do teste $U_{p,q}$ é dado a seguir. No capítulo 2, no ítem 2.7, consideramos a discriminação entre a Iris Versicolor e a Iris Setosa, baseados no comprimento e na largura dos caules. Comentamos então que o comprimento e a largura das pétalas eram variáveis altamente discriminantes. Mostremos agora que a discriminação baseada nas 4 medidas é significantemente superior à obtida apenas com as medidas do caule.

No sub-ítem anterior (3.9.1) calculamos D^2_4 e obtivemos

$$D^2_4 = 103,2119.$$

Calculemos agora D_2^2 isto é o D^2 de Mahalanobis baseado no comprimento e largura dos caules. De (2.7) temos que

$$y = 0,1167x_1 - 0,1443x_2$$

sendo $b_1 = 0,1167$ e $b_2 = -0,1443$,

mas

$$D_2^2 = \sum_{i=1}^{2} \sum_{j=1}^{2} w^{ij} d_i d_j = \sum_{i=1}^{2} \sum_{j=1}^{2} (n_1+n_2-2) S^{ij} d_i d_j =$$

$$= (n_1+n_2-2) \sum_{i=1}^{2} b_i d_i$$

$$D_2 = 98\{0,1167\times 0,930 - 0,1443\times(-0,658)\} = 98\times 0,2034 = 19.9410$$

portanto

$$U_{p,q} = \frac{95\times 50\times 50\times(103,2119-19,9410)}{2(100\times 98+50\times 50\times 19,9410)} = 165,767$$

que comparado com o F de 2 e 95 graus de liberdade ao nível de 1% mostra que a melhoria da discriminação com o incremento das medidas das pétalas é altamente significativa como era de se esperar (o comprimento da pétala sózinho permite uma discriminação a primeira vista!).

3.9.3 - OUTROS TESTES

Outros testes ligados ao problema de discriminação podem ser encontrados em Rao (referências 9 e 10).

Um deles proposto por Fisher se destina a testar a

validade de uma função linear dada, como função discriminante entre duas populações. A construção de tais funções lineares é normalmente motivada por razões intuitivas a partir da simples observação das medidas.

Um outro se destina a testar se o quociente entre dois coeficientes b_i e b_j da função discriminante assume determinado valor. (Sabemos que os coeficientes da função discriminante não são únicos mas que estão dois a dois ligados por uma razão constante).

- □ -

CAPÍTULO 4

DISCRIMINAÇÃO ENTRE k POPULAÇÕES

4.1 - INTRODUÇÃO

Nosso objetivo inicial é a determinação de uma regra discriminatória ótima no sentido de minimizar a perda esperada (risco) no processo de classificação. Para tal particionaremos o espaço amostral em k regiões $R_1,\ldots,R_k$ de forma que se um elemento pertencer a R_i ele será classificado em π_i. Para a determinação da partição $R = (R_1,\ldots,R_k)$ acima descrita consideraremos dois casos: probabilidades a priori conhecidas e probabilidades a priori desconhecidas.

Aplicaremos os resultados obtidos para o caso de k populações normais com mesma matriz de covariância. O caso em que os parâmetros populacionais são desconhecidos também é abordado.

Um novo critério de classificação é introduzido com a construção da "região de dúvida" de Rao. Os elementos pertencentes à região de dúvida não são classificados pois as probabilidades de erro de classificação para estes elementos superam limites pré-estabelecidos.

Dois testes ligados ao processo de discriminação são apresentados. Um deles constitui base teórica para o programa BMD07M, da Biomedical Computer Programs, utilizado na escolha conveniente das variáveis discriminadoras. A apresentação deste programa encerra o Capítulo.

Será mantida a notação do Capítulo anterior com a devida extensão dos índices.

4.2 - DETERMINAÇÃO DE R QUANDO AS PROBABILIDADES A PRIORI SÃO CONHECIDAS

Se conhecemos as probabilidades a priori p_i ($p_1+\ldots+p_k=1$), a perda esperada no processo discriminatório ao adotar a regra (partição) R será calculada pela expressão

$$r(R) = p_1 r(1,R) + \ldots + p_k r(k,R)$$

onde a perda esperada ao classificar um elemento de π_i, $r(i,R)$, é:

$$r(i,R) = \sum_{\substack{j=1 \\ j\neq i}}^{k} C(j|i)P(j|i,R) = \sum_{\substack{j=1 \\ j\neq i}}^{k} \int_{R_j} C(j|i)f_i(\underline{x})d\underline{x}$$

Assim, a perda r(R) pode ser reescrita na forma

$$r(R) = \sum_{i=1}^{k} p_i \sum_{\substack{j=1 \\ j\neq i}}^{k} \int_{R_j} C(j|i)f_i(\underline{x})d\underline{x} = \sum_{j=1}^{k} \int_{R_j} \{ \sum_{\substack{i=1 \\ i\neq j}}^{k} p_i C(j|i)f_i(\underline{x})\}d\underline{x}$$

ou simplesmente:

$$r(R) = \sum_{j=1}^{k} \int_{R_j} F_j(\underline{x})d\underline{x} \text{ onde } F_j(\underline{x}) = \sum_{\substack{i=1 \\ i \neq j}}^{k} p_i C(j|i) f_i(\underline{x})$$

Pelo teorema demonstrado no Apêndice C, decorre que a regra R que minimiza r(R) é assim definida

$$R_1: F_1(\underline{x}) \leq F_2(\underline{x}), \ldots, F_1(\underline{x}) \leq F_k(\underline{x})$$

$$\ldots \quad \ldots\ldots \quad \ldots\ldots \quad \ldots\ldots \quad \ldots\ldots \quad \ldots\ldots$$

$$R_k: F_k(\underline{x}) \leq F_1(\underline{x}), \ldots, F_k(\underline{x}) \leq F_{k-1}(\underline{x})$$

Se o vetor $\underline{x}$ das medidas associadas ao indivíduo E satisfizer simultaneamente à definição de duas ou mais regiões é indiferente classificar E em uma qualquer das populações associadas a estas regiões. A menos destes vetores $\underline{x}$ a partição R é única.

Para melhor visualizar a definição da partição acima consideremos alguns casos particulares. Lembramos que conforme foi visto anteriormente consideramos $C(i|i) = 0$ e $C(i|j) > 0$, $i \neq j$.

4.2.1 - <u>k = 2</u>

Neste caso teremos: $R = (R_1, R_2)$

$$F_1(\underline{x}) = p_2 C(1|2) f_2(\underline{x})$$

$$F_2(\underline{x}) = p_1 C(2|1) f_1(\underline{x})$$

e as regiões R_1, R_2 ficam definidas por:

$$R_1: p_2C(1|2)f_2(\underline{x}) \leq p_1C(2|1)f_1(\underline{x})$$

$$R_2: p_1C(2|1)f_1(\underline{x}) \leq p_2C(1|2)f_2(\underline{x})$$

coincidindo evidentemente com os resultados obtidos no Capítulo anterior.

4.2.2 - k = 3

Neste caso teremos: $R = (R_1, R_2, R_3)$

$$F_1(\underline{x}) = p_2C(1|2)f_2(\underline{x}) + p_3C(1|3)f_3(\underline{x})$$

$$F_2(\underline{x}) = p_1C(2|1)f_1(\underline{x}) + p_3C(2|3)f_3(\underline{x})$$

$$F_3(\underline{x}) = p_1C(3|1)f_1(\underline{x}) + p_2C(3|2)f_2(\underline{x})$$

e se os custos $C(i|j)$ $i \neq j$, forem iguais, as regiões serão

$$R_1: p_1f_1(\underline{x}) \geq p_2f_2(\underline{x}) \quad , \quad p_1f_1(\underline{x}) \geq p_3f_3(\underline{x})$$

$$R_2: p_2f_2(\underline{x}) \geq p_1f_1(\underline{x}) \quad , \quad p_2f_2(\underline{x}) \geq p_3f_3(\underline{x})$$

$$R_3: p_3f_3(\underline{x}) \geq p_1f_1(\underline{x}) \quad , \quad p_3f_3(\underline{x}) \geq p_2f_2(\underline{x})$$

o que equivale a dizer que um elemento será classificado na população em relação à qual sua probabilidade a posteriori foi maior.

4.3 - DETERMINAÇÃO DE R QUANDO AS PROBABILIDADES A PRIORI SÃO DESCONHECIDAS

No caso em que as probabilidades a priori não são

conhecidas e nem mesmo estimáveis, optaremos teoricamente pela determinação da regra R minimax. Isto significa que consideraremos uma regra R ótima se minimizar o maior dos riscos $r(i,R)$ (perda esperada ao classificar um elemento de π_i).

Sendo o conjunto das regras Bayesianas sob certas condições (vide Apêndice B) uma classe completa mínima, buscaremos dentre as regras desta classe a solução minimax. A solução existe e será adotada como critério discriminante.

Como já vimos, uma regra Bayesiana associada a um vetor de probabilidades a priori $\underline{p} = (p_1, \ldots, p_k)$ é uma regra que minimiza a perda esperada no processo discriminatório, isto é, minimiza $r(R)$,

$$r(R) = p_1 r(1,R) + \ldots + p_k r(k,R)$$

Determinar a solução minimax significa obter (por tentativas) um vetor $\underline{p}^*$ tal que a regra Bayesiana a ele associada defina o critério discriminante que minimize o maior dos riscos condicionais $r(i,R)$, $i=1,\ldots,k$. Isto levará a uma R^* tal que $r(1,R^*) = \ldots = r(k,R^*)$.

Omitiremos a demonstração da existência da solução minimax, que pode ser encontrada em Wald, A. (1950) — Statistical Decision Functions -Wiley - N. York. A verificação de que R^* é solução minimax é simples: se existisse uma regra S tal que $\max\{r(i,S),\ i=1,\ldots,k\} < r(1,R^*) = \ldots = r(k,R^*)$ isto negaria o fato de que a regra Bayesiana R^* é admissível.

Em termos de aplicação é fácil perceber o trabalho

absurdo que se terá na determinação do vetor $\underline{p}^*$ especialmente no caso de mais de duas populações, aumentando barbaramente com o aumento de k. Rao em 1948 iniciou um trabalho de pesquisa sobre aproximações aplicáveis na determinação de regiões minimax quando $k \geq 3$.

Embora nos pareça a melhor opção teórica para o problema, cremos que em termos práticos é mais conveniente optar pela proposição de Rao para o caso de probabilidades a priori desconhecidas: adotar $p_1 = \ldots = p_k$.

4.4 - DISCRIMINAÇÃO ENTRE k POPULAÇÕES NORMAIS

Neste ítem estenderemos os resultados obtidos em (4.2) e (4.3) para o caso particular em que as k populações são normais multivariadas com mesma matriz de covariância.

Se $\underline{X} = (X_1, \ldots, X_p)$, sob π_i teremos $\underline{X} \sim N_p(\underline{\mu}_i, \Sigma)$, isto é

$$f_i(\underline{x}) = \frac{1}{(2\pi)^{p/2} |\Sigma|^{1/2}} \exp\{ -\frac{1}{2}(\underline{x}-\underline{\mu}_i)' \Sigma^{-1} (\underline{x}-\underline{\mu}_i)\}$$

Suporemos os custos $C(i|j)$, $i \neq j$, iguais.

4.4.1 - Probabilidades a priori conhecidas

Estendendo para k qualquer os resultados obtidos em (4.2.2), as regiões $R_1, \ldots, R_k$ ficam definidas por

$$R_i : p_i f_i(\underline{x}) \geq p_j f_j(\underline{x}) \quad j=1,\ldots,k \; ; \; j \neq i$$

Sendo o logaritmo neperiano (log) uma função crescente, podemos escrever a definição de R_i da seguinte forma

$$R_i: \log f_i(\underline{x}) - \log f_j(\underline{x}) \geq \log p_j - \log p_i \quad j=1,\ldots,k;\ j\neq i$$

Denotando $\log f_i(\underline{x}) - \log f_j(\underline{x})$ por u_{ij} teremos

$$R_i: u_{ij} \geq \log p_j - \log p_i \quad j=1,\ldots,k;\ j\neq i$$

No caso particular em que $p_1 = \ldots = p_k$, resulta

$$R_i: u_{ij} \geq 0 \quad j=1,\ldots,k;\ j\neq i$$

É conveniente notar que $U_{ij} = -U_{ji}$ e que ao classificar um indivíduo devemos calcular apenas (k-1) funções U_{ij} convenientemente escolhidas e não k(k-1)/2 como poderia parecer.

4.4.2 - Probabilidades a priori desconhecidas

Conforme vimos em (4.3), no caso das probabilidades a priori não serem conhecidas optaremos em princípio pelo método minimax, que determinará uma partição R em que

$$r(1,R) = \ldots = r(k,R).$$

Comentamos também a dificuldade prática que surge na determinação de R, motivo pelo qual sugerimos a solução proposta por Rao que consiste em adotar $p_1 = \ldots = p_k$. Neste ítem apresentaremos as equações a partir das quais se determina a solução (R) minimax, bem como apresentaremos a definição do critério discriminante a partir da proposição de Rao.

Sabemos que

$$r(i,R) = \sum_{\substack{j=1 \\ j\neq i}}^{k} C(j|i)P(j|i,R)$$

Supondo $C(i|j) = C$, $i\neq j$, onde C é uma constante positiva, teremos

$$r(i,R) = C \sum_{\substack{j=1 \\ j\neq i}}^{k} P(j|i,R)$$

$$= C\{1 - P(i|i,R)\}$$

do que decorre que a solução minimax corresponde à partição R, tal que

$$P(1|1,R) = \ldots = P(k|k,R)$$

onde $P(i|i,R)$ representa a probabilidade de classificação correta de um elemento de π_i.

Como as regiões $R_1,\ldots,R_k$ da solução Bayesiana R são definidas por

$$R_i: u_{ij} \geq q_j - q_i \quad j=1,\ldots,k;\ j\neq i$$

onde $q_j = \log p_j$, $j=1,\ldots,k$, isto significa que

$$P(i|i,R) = P(U_{i1} \geq q_1 - q_i, \ldots, U_{i,i-1} \geq q_{i-1} - q_i, U_{i,i+1} \geq$$
$$\geq q_{i+1} - q_i, \ldots, U_{ik} \geq q_k - q_i)$$

ou

$$P(i|i,R) = \int_{q_k - q_i}^{\infty} \ldots \int_{q_1 - q_i}^{\infty} f_i \, du_{i1} \ldots du_{i,i-1} du_{i,i+1} \ldots du_{ik}$$

onde f_i é a função densidade de probabilidade da variável aleatória (k-1)-dimensional $\underline{U}_i = (U_{i1},\ldots,U_{i,i-1},U_{i,i+1},\ldots,U_{ik})$. Como a variável $\underline{X} = (X_1,\ldots,X_p)$ tem distribuição normal, a variável

$$U_{ij} = \underline{X}'\Sigma^{-1}(\underline{\mu}_i-\underline{\mu}_j) - \frac{1}{2}(\underline{\mu}_i+\underline{\mu}_j)'\Sigma^{-1}(\underline{\mu}_i-\underline{\mu}_j)$$

sendo combinação linear de $X_1,\ldots,X_p$ terá distribuição normal. A variável $\underline{U}_i$ será normal pois toda combinação linear de suas componentes será uma combinação linear de X_1, $X_2,\ldots,X_p$ e portanto normal.

Sob a hipótese $\underline{X}\in\pi_i$ então a média da variável U_{ij} será dada por

$$E_i(U_{ij}) = \frac{1}{2}(\underline{\mu}_i-\underline{\mu}_j)'\Sigma^{-1}(\underline{\mu}_i-\underline{\mu}_j)$$

e a variância por

$$Var_i(U_{ij}) = (\underline{\mu}_i-\underline{\mu}_j)'\Sigma^{-1}(\underline{\mu}_i-\underline{\mu}_j)$$

notando-se que $(\underline{\mu}_i-\underline{\mu}_j)'\Sigma^{-1}(\underline{\mu}_i-\underline{\mu}_j)$ é a distância generalizada de Mahalanobis entre π_i e π_j.

A covariância entre U_{ij} e U_{im} será dada por

$$Cov_i(U_{ij},U_{im}) = (\underline{\mu}_i-\underline{\mu}_j)'\Sigma^{-1}(\underline{\mu}_i-\underline{\mu}_m)$$

Para a obtenção da solução minimax devemos obter as constantes $q_1,\ldots,q_k$ (por um processo iterativo) para as quais

$$P(1|1,R) = \ldots = P(k|k,R),$$

isto é, para que as integrais acima sejam iguais para $i=1,\ldots,k$.

Se não quizermos adotar a solução minimax, que como observamos pelo acima exposto não é nada simples de obter, optaremos pela sugestão de Rao, isto é, $p_1=\ldots=p_k$ com o que as regiões $R_1,\ldots,R_k$ ficam simplesmente definidas por

$$R_i: u_{ij} \geq 0 \quad j=1,\ldots,k;\ i\neq j.$$

4.4.3 - Parâmetros Populacionais Desconhecidos

Temos até o momento admitido o conhecimento dos parâmetros populacionais $\underline{\mu}_i$ e Σ. Se não os conhecermos no entanto, usaremos suas estimativas se bem que com isto estaremos aumentando a probabilidade de erro no processo de classificação. Veremos no entanto que se as amostras forem suficientemente grandes, o erro introduzido pela utilização das estimativas pode ser desprezado.

A estimativa $\bar{\underline{x}}_i$ de $\underline{\mu}_i$ é definida por

$$\bar{\underline{x}}_i = \frac{1}{n_i} \sum_{m=1}^{n_i} \underline{x}_{im}$$

onde $\underline{x}_{im}$ é o vetor das medidas do m-ésimo elemento da amostra de π_i.

A estimativa W de Σ é definida por

$$\left(\sum_{k=1}^{k} n_i-k\right)W = \sum_{i=1}^{k} \sum_{m=1}^{n_i} (\underline{x}_{im}-\bar{\underline{x}}_i)(\underline{x}_{im}-\bar{\underline{x}}_i)'$$

E a estatística V_{ij} correspondente a U_{ij} será dada por

$$V_{ij} = \underline{X}'W^{-1}(\bar{\underline{X}}_i - \bar{\underline{X}}_j) - \frac{1}{2}(\bar{\underline{X}}_i + \bar{\underline{X}}_j)'W^{-1}(\bar{\underline{X}}_i - \bar{\underline{X}}_j)$$

Como $\bar{\underline{X}}_i$ converge em probabilidade para $\underline{\mu}_i$ e W para Σ, a distribuição limite de V_{ij} será U_{ij}. Isto é garantido pelo teorema de Slutsky (cf. H. CRAMER - "Mathematical Methods of Statistics") sobre a convergência em probabilidade de funções racionais de variáveis aleatórias que convergem em probabilidade a determinadas constantes. Isto, traduzido em termos práticos significa, como comentamos acima, que para amostras suficientemente grandes o erro devido à utilização de estimativas pode ser desprezado.

4.5 - A REGIÃO DE DÚVIDA

Neste ítem apresentamos a construção da região de dúvida, R_0, introduzida na Análise Discriminante por Rao. O aparecimento de R_0 no processo discriminante é consequência de uma limitação pré-estabelecida da probabilidade de erro de classificação.

Definidas as regiões R_1, R_2, ..., R_k conforme visto nos ítens anteriores, a probabilidade de que um elemento seja corretamente classificado em π_i é dada por

$$\beta_1 = p_1 P(1|1,R) = \int_{R_1} p_i f_1(\underline{x})d\underline{x}$$

enquanto que a probabilidade de que um elemento seja classificado erradamente em π_1 é

$$\alpha_1 = p_2 P(1|2,R) + \ldots + p_k P(1|k,R)$$

Ocorrem situações em que desejamos limitar o valor α_1 por um máximo α_1^0. Se a região R_1 for tal que $\alpha_1 \leq \alpha_1^0$ então um elemento que pertença a R_1 será classificado em π_1 dentro do limite de probabilidade de erro (α_1^0) previamente estabelecido. Caso contrário, isto é, se $\alpha_1 > \alpha_1^0$ devemos determinar uma região C_1 tal que a probabilidade de classificar erradamente em π_1 um elemento dessa região seja menor ou no máximo igual a α_1^0. À região $D_1 = R_1 - C_1$ denominaremos "região de dúvida associada a π_1". Os elementos pertencentes a D_1 não serão classificados embora sua pertinência a π_1 seja provável.

É claro que dentre as regiões C_1 que satisfazem esse requisito, devemos escolher aquela para a qual a probabilidade de classificação correta em π_1 for máxima. Em outras palavras, procuramos a região C_1 para a qual

$$\int_{C_1} p_1 f_1(\underline{x}) d\underline{x} \tag{1}$$

é máxima, sujeita à restrição

$$\int_{C_1} p_2 f_2(\underline{x}) d\underline{x} + \ldots + \int_{C_1} p_k f_k(\underline{x}) d\underline{x} \leq \alpha_1^0 \tag{2}$$

A determinação desta região decorre de imediato da aplicação do lema de Neyman-Pearson (Apêndice D). A região C_1 sujeita a restrição (2) que maximiza (1) é dada por

$$C_1: p_1 f_1(\underline{x}) \geq A(p_2 f_2(\underline{x}) + \ldots + p_k f_k(\underline{x}))$$

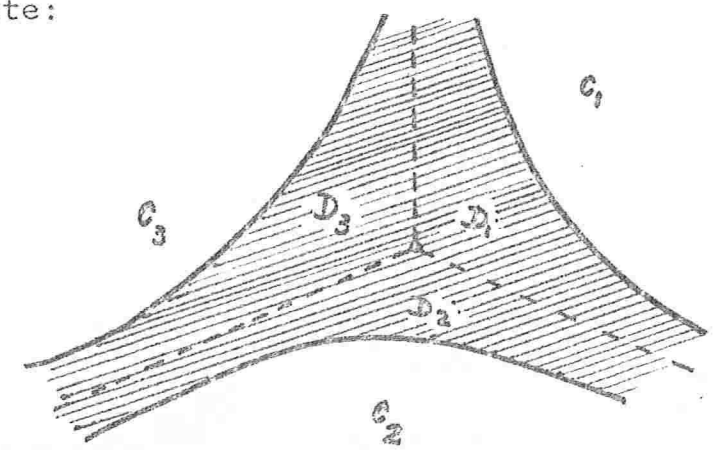
onde A é uma constante convenientemente escolhida, isto é, A deve ser tal que

$$\int_{C_1} \{p_2 f_2(\underline{x}) + \ldots + p_k f_k(\underline{x})\} d\underline{x} = \alpha_1^0$$

Obtida a região C_1, a região $D_1 = R_1 - C_1$ está definida.

As regiões $C_2, \ldots, C_k$ correspondentes às demais populações são obtidas de forma análoga e consequentemente as regiões $D_2, \ldots, D_k$. À união das regiões $D_1, D_2, \ldots, D_k$ denominaremos "região de dúvida de Rao" e será denotada por R_0.

Para o caso de três populações teremos esquematicamente:



onde a parte hachurada representa R_0.

É importante notar que na determinação de R_0 utilizamos as probabilidades a priori $p_1,\ldots,p_k$. No caso destas não serem conhecidas ou nem mesmo estimáveis poderemos adotar $p_1 = \ldots = p_k$.

Ao invés de limitar a probabilidade de erro ao classificar um elemento em π_i, como fizemos acima, pode-se desejar limitar a probabilidade de erro ao classificar um elemento de π_j em π_i, $i \neq j$. Este segundo caso difere evidentemente do primeiro, mas a determinação da região de dúvida é similar sendo apresentada por Rao (referência 9) num desenvolvimento independente da consideração das probabilidades a priori.

Rao sugere que se considere o primeiro tipo de limitação no caso em que as probabilidades a priori são conhecidas e o segundo tipo caso contrário. Quer nos parecer no entanto que a construção das regiões de dúvida não deve se basear no conhecimento ou não de informação a priori e sim no tipo de erro que desejamos limitar.

4.6 - DOIS TESTES RELACIONADOS COM A DISCRIMINAÇÃO ENTRE k POPULAÇÕES

Apresentaremos neste ítem dois testes intimamente ligados ao problema de discriminação. Admitiremos que as populações são normais multivariadas com mesma matriz de cova

riância.

4.6.1 - Teste da Igualdade de Médias

Este teste deve evidentemente ser aplicado antes de iniciar a procura de uma regra discriminante entre as k populações. A razão disto é simples: a discriminação só terá sentido se não houver coincidência de populações. O teste que iremos expor (cf. Rao, referência 10, 8d.4) é uma extensão da Análise de Variância para o caso de populações multivariadas, sendo conhecido por Análise de Dispersão.

Se $\underline{X} = (X_1,\ldots,X_p)$ é o vetor das variáveis aleatórias discriminantes, para a população π_r, $r=1,\ldots,k$, teremos $\underline{X} \sim N_p(\underline{\mu}_r,\Sigma)$. Testaremos a hipótese H_0: $\underline{\mu}_1 = \ldots = \underline{\mu}_k$ contra a hipótese alternativa da não igualdade entre as k médias.

Sejam $n_1,\ldots,n_k$ os tamanhos das amostras correspondentes às populações $\pi_1,\ldots,\pi_k$ e representemos por n a soma $n_1+\ldots+n_k$. Consideremos a seguinte notação:

$$\bar{x}_{ri} = \frac{1}{n_r}\sum_{m=1}^{n_r} x_{rim} \qquad \bar{x}_i = \frac{1}{n}\sum_{r=1}^{k} n_r\,\bar{x}_{ri}$$

$B = (B_{ij})$: matriz da soma de produtos entre grupos, onde

$$B_{ij} = \sum_{r=1}^{k} n_r(\bar{x}_{ri}-\bar{x}_i)(\bar{x}_{rj}-\bar{x}_j)$$

$W = (W_{ij})$: matriz da soma de produtos dentro, onde:

$$W_{ij} = \sum_{r=1}^{k} \sum_{m=1}^{n_r} (x_{rim}-\bar{x}_{ri})(x_{rjm}-\bar{x}_{rj})$$

$S = (S_{ij})$: matriz da soma de produtos total, onde

$$S_{ij} = \sum_{r=1}^{k} \sum_{m=1}^{n_r} (x_{rim}-\bar{x}_{i})(x_{rjm}-\bar{x}_{j})$$

sendo que das expressões acima podemos concluir que

$$B + W = S$$

O teste baseia-se no critério de Wilks, Λ, onde

$$\Lambda = \frac{|W|}{|W+B|}$$

A estatística Λ se distribui como o produto de p variáveis aleatórias independentes com distribuição Beta. Em alguns casos, a utilização de Λ sob certas transformações nos leva a distribuições F exatas. Em Rao (referência 10) encontramos um resumo de tais transformações, especificamente para os casos em que $k = 2$, $k = 3$, $p = 1$, $p = 2$. Para outros valores de p e k podemos utilizar o teste aproximado proposto por Rao (1951) que se resume no uso da estatística

$$\frac{1 - \Lambda^{1/s}}{\Lambda^{1/s}} \quad \frac{ms - 2h}{p(k-1)}$$

como uma F com $p(k-1)$ e $(ms-2h)$ graus de liberdade, onde

$$m = n-1-\frac{p+k}{2}; \quad s = \sqrt{\frac{p^2(k-1)^2-4}{p^2+(k-1)^2-5}}; \quad h = \frac{p(k-1)-2}{4}$$

onde (ms-2h) não é necessariamente inteiro.

Outro teste aproximado, baseado no critério de Wilks, foi proposto por Bartlett (1947) e se resume ao uso da estatística

$$-m \log_e \Lambda$$

como uma χ^2 com $p(k-1)$ graus de liberdade.

4.6.2 - Teste da Melhoria de Discriminação

Vamos supor que a discriminação entre k populações esteja baseada em p variáveis aleatórias $X_1,\ldots,X_p$, e desejamos testar se a utilização de q variáveis adicionais $X_{p+1}, \ldots, X_{p+q}$ introduz uma melhoria significativa na discriminação entre essas k populações. Para esse fim, apresentamos aqui um teste (cf. Rao, referência 10, 8 c.4) que como o anterior se baseia no critério Λ de Wilks. Este teste é a base teórica do processo stepwise BMD07M que comentaremos no ítem seguinte.

É importante notar que não testaremos a significância da discriminação baseada nas q variáveis adicionais (para o que usaríamos o teste anterior), mas sim se a discriminação (adicional) introduzida por essas variáveis, quando o efeito das p variáveis iniciais é fixado, é significativa.

As matrizes W e S (definidas no teste anterior) para o conjunto das p+q variáveis podem ser particionadas da

seguinte forma:

$$W = \begin{pmatrix} W_{pp} & W_{pq} \\ W_{qp} & W_{qq} \end{pmatrix} \qquad S = \begin{pmatrix} S_{pp} & S_{pq} \\ S_{qp} & S_{qq} \end{pmatrix}$$

onde os índices p e q se referem respectivamente aos conjuntos das p variáveis iniciais e das q variáveis adicionais.

As matrizes das somas de produtos das q variáveis adicionais, quando o efeito das p variáveis iniciais é fixado, são dadas por

$$W(q|p) = W_{qq} - W_{qp}\, W_{pp}^{-1}\, W_{pq}$$

e

$$S(q|p) = S_{qq} - S_{qp}\, S_{pp}^{-1}\, S_{pq}$$

Sob a hipótese de discriminação adicional nula, o quociente dos determinantes

$$\frac{|W(q|p)|}{|S(q|p)|}$$

se distribui segundo a Λ de Wilks.

Para efetuar o teste nos casos em que $q \neq 1$, $q \neq 2$, $k \neq 2$, $k \neq 3$ (quando a utilização de certas transformações sobre Λ nos leva a testes F exatos) podemos utilizar o teste aproximado sugerido por Rao (1951) que consiste em usar a estatística

$$\frac{1-\Lambda^{1/s}}{\Lambda^{1/s}} \quad \frac{ms - 2h}{q(k-1)}$$

como uma F com q(k-1) e (ms-2h) graus de liberdade, onde

$$m = n-p-1-\frac{q+k}{2}; \quad s = \sqrt{\frac{q^2(k-1)^2-4}{q^2+(k-1)^2-5}}; \quad h = \frac{q(k-1)-2}{4}$$

No caso particular q = 1, isto é, quando queremos testar a melhoria da discriminação introduzida por uma variável adicional, Λ será o quociente dos escalares $W(q|p)$ e $S(q|p)$ e a estatística

$$\frac{1-\Lambda}{\Lambda} \quad \frac{n-p-k}{k-1}$$

tem distribuição F com (k-1) e (n-p-k) graus de liberdade.

4.7 - A ESCOLHA CONVENIENTE DE VARIÁVEIS

Uma das dificuldades que surgem no estudo da discriminação entre populações é a seleção das variáveis discriminadoras. Se conhecermos um número "p" de variáveis discriminadoras, podemos estar interessados (normalmente por motivos de custo operacional) em selecionar dentre estas, o menor conjunto de q variáveis ($q \leq p$) de forma tal que a discriminação baseada nessas q variáveis seja significativa e que a consideração de quaisquer das p - q variáveis restantes não introduza melhoria significativa na discriminação baseada nas q variáveis selecionadas.

Um dos processos utilizados na seleção das variáveis é o Stepwise Discriminant Analysis (BMD07M) da Biomedi

cal Computer Programs do qual apresentaremos um resumo do procedimento.

Inicialmente, selecionamos dentre as p variáveis aquela que "melhor discrimina" entre as populações dadas. Isto é medido pela F de uma análise de variância a um fator controlado. A variável à qual corresponder maior F será selecionada (entered) desde que este F supere um mínimo pré-estabelecido (F to include).

Nos passos seguintes consideramos as "melhorias de discriminação" introduzidas por cada uma das variáveis ainda não selecionadas, em relação à discriminação devida às variáveis já escolhidas. Será selecionada como nova variável aquela que apresentar "maior melhoria de discriminação" o que é medido pela F de um teste de melhoria de discriminação. Esta F deverá ser maior que o "F to include", caso contrário o processo para de selecionar.

Um ponto importante a destacar é que o fato de uma variável ter sido escolhida em um determinado passo,não significa que ela será mantida no conjunto final das q variáveis selecionadas pelo processo. A cada passo (a partir da escolha da segunda variável) são calculados para cada uma das ℓ variáveis escolhidas o "F to remove" e se este for menor que um limite pré-estabelecido (F to exclude poderá ser eliminada (é eliminada a variável que apresentar o menor "F to remove" desde que este seja menor que o "F to exclude"). O "F to remove"é o F obtido no teste de melhoria de discri-

minação devido à variável considerada em relação à discriminação devida às (ℓ-1) variáveis escolhidas restantes.

O processo para quando nenhuma das variáveis a serem selecionadas apresentar melhoria significativa na discriminação ou quando todas as variáveis já tiverem sido selecionadas e nenhuma deva ser eliminada.

- □ -

CAPÍTULO 5

EXEMPLO DE APLICAÇÃO

5.1 - O PROBLEMA

Consideremos duas populações assim definidas: a primeira (π_1) constituída por indivíduos que apresentam distúrbios de tireóide e a segunda (π_2) por indivíduos sem problemas de tireóide (indivíduos normais). Estudaremos a discriminação entre essas duas populações com base nas medidas de Glicose (X_1) Insulina (X_2) e Nefa (X_3) efetuadas para cada um dos elementos dessas populações.

Admitiremos a normalidade das distribuições bem como a igualdade de dispersão nas duas populações.

O presente exemplo é apresentado apenas com a finalidade de ilustrar como abordar um problema de discriminação, com a utilização de técnicas vistas nos Capítulos anteriores. Os dados considerados são valores fictícios.

5.2 - APRESENTAÇÃO DOS DADOS

Os dados apresentados na página 85 correspondem aos elementos de duas amostras aleatórias extraídas das popula-

ções em questão, cada uma com 20 elementos ($n_1 = n_2 = 20$)

AMOSTRA DE π_1			AMOSTRA DE π_2		
X_1	X_2	X_3	X_1	X_2	X_3
72	18	808	79	4	690
76	17	854	78	7	680
73	20	860	79	8	673
75	19	864	80	8	652
76	21	871	83	7	669
78	20	638	75	9	681
80	19	693	77	7	653
80	18	742	74	9	686
78	17	780	74	10	686
82	18	878	74	5	639
78	22	763	75	9	614
75	20	812	74	12	600
77	22	822	75	11	622
76	20	857	75	14	636
79	19	881	74	10	645
78	19	659	72	9	654
78	15	687	75	8	636
82	14	680	73	8	617
79	15	694	74	7	618
76	19	756	75	11	635

Dos dados acima, obtemos os seguintes valores:

MÉDIAS AMOSTRAIS

	π_1	π_2	DIFERENÇAS $(\pi_1-\pi_2)$
X_1	$\bar{x}_{11}$ = 77,40	$\bar{x}_{21}$ = 75,75	d_1 = 1,65
X_2	$\bar{x}_{12}$ = 18,60	$\bar{x}_{22}$ = 8,65	d_2 = 9,95
X_3	$\bar{x}_{13}$ =779,95	$\bar{x}_{23}$ =649,30	d_3 = 130,65

MATRIZ DE COVARIÂNCIA AMOSTRAL

$$W = \begin{pmatrix} 7,1723 & -2,3302 & -27,2920 \\ -2,3302 & 5,0881 & 21,8341 \\ -27,2920 & 21,8341 & 3.728,0210 \end{pmatrix}$$

à qual corresponde a matriz inversa

$$W^{-1} = \begin{pmatrix} 0,1660 & 0,0726 & 0,0007 \\ 0,0726 & 0,2333 & -0,0008 \\ 0,0007 & -0,0008 & 0,0002 \end{pmatrix}$$

5.3 - TESTE DA SIGNIFICÂNCIA DA DISCRIMINAÇÃO

O primeiro passo a ser dado é verificar a significância da discriminação baseada nas variáveis Glicose (X_1), Insulina (X_2) e Nefa (X_3). Para isto calculamos a estatística

$$F = \frac{(n_1+n_2-p-1)}{(n_1+n_2-2)p} \frac{n_1 n_2}{n_1+n_2} D^2$$

que sob a hipótese de coincidência das duas populações (não

significância da discriminação) tem distribuição F com p e n_1+n_2-p-1 graus de liberdade.

Como

$n_1 = n_2 = 20$ tamanhos das amostras

$p = 3$ número de variáveis

$D^2 = \sum_{i=1}^{3} \sum_{j=1}^{3} w^{ij} d_i d_j = 27,5654$ distância de Mahalanobis

obtemos $F = 87,0486$. Sendo $F_{3,36}$ (5%) = 2,86 concluímos que a discriminação entre π_1 e π_2 baseada nas variáveis Glicose, Insulina e Nefa é significativa.

5.4 - A DISCRIMINAÇÃO

Como não conhecemos os parâmetros populacionais $\underline{\mu}_i$ e Σ, utilizaremos seus estimadores $\bar{\underline{X}}_i$ e W e portanto as regiões R_1 e R_2 ficarão definidas com auxílio da estatística V

$$R_1: v \geq \log k \text{ e } R_2: v \leq \log k$$

onde $k = \dfrac{C(1|2)p_2}{C(2|1)p_1}$.

Admitindo os custos C(1|2) e C(2|1) iguais resulta

$$\log k = \log p_2 - \log p_1$$

As probabilidades p_1 e p_2 são desconhecidas, o que nos leva a utilizar o processo minimax para determinar R_1 e R_2. No entanto, como observamos no Capítulo 3, para o caso

particular em que as duas populações são normais com mesma matriz de covariância, igualados os custos $C(1|2)$ e $C(2|1)$, a solução minimax equivale à solução obtida ao adotar $p_1 = p_2$. Em outras palavras, as regiões R_1 e R_2 serão definidas por

$$\boxed{R_1 : v \geq 0} \quad \text{e} \quad \boxed{R_2 : v \leq 0.}$$

Sendo

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

teremos, de acordo com os dados apresentados:

$$\bar{\underline{x}}_1 = \begin{pmatrix} 77,40 \\ 18,60 \\ 779,95 \end{pmatrix}, \quad \bar{\underline{x}}_2 = \begin{pmatrix} 75,75 \\ 8,65 \\ 649,30 \end{pmatrix}$$

e portanto

$$v = 1,0877x_1 + 2,3366x_2 + 0,0193x_3 - 128,9191$$

Podemos então escrever

$$R_1 : 1,0877x_1 + 2,3366x_2 + 0,0193x_3 \geq 128,9191$$

$$R_2 : 1,0877x_1 + 2,3366x_2 + 0,0193x_3 \leq 128,9191$$

sendo interessante notar que o primeiro membro das desigualdades acima é a função discriminante de Fisher (Capítulo 2).

A utilização da regra discriminante acima é bastante simples. Se por exemplo, um indivíduo (E) apresenta as seguintes medidas: Glicose: 75, Insulina: 10, Nefa: 700, então para ele teremos:

$$y = 1,0877x_1 + 2,3366x_2 + 0,0193x_3 = 118,4535$$

81,5775 + 23,366 +

ou

$$v = y - 128,9191 = -10,4656$$

pois o seu "v" é < que 0

o que indica que esse elemento pertence a R_2 e deve ser, portanto, classificado em π_2, isto é, é um indivíduo normal.

* * *

Vamos supor que consultando manuais médicos concluímos que a proporção entre indivíduos com distúrbios de tireóide e indivíduos normais é de 1:50 (valores fictícios). Isto quer dizer que $p_1 = (1/51) = 0,0196$ e que $p_2 = (50/51) = 0,9804$.

Como $\log p_1 = -3,9318$ e $\log p_2 = -0,0198$ teremos $\log k = \log p_2 - \log p_1 = 3,9120$ e as regiões R_1 e R_2 ficam definidas por

$$R_1 : v \geq 3,9120 \quad \text{e} \quad R_2 : v \leq 3,9120$$

ou

$$R_1 : 1,0877x_1 + 2,3366x_2 + 0,0193x_3 \geqslant 132,8311$$

$$R_2 : 1,0877x_1 + 2,3366x_2 + 0,0193x_3 \leqslant 132,8311$$

Comparando estas regiões com as obtidas no caso anterior verificamos que o ponto de separação entre as duas populações se deslocou em sentido de π_1 (128,9191 → 132,8311) o que é intuitivamente lógico, visto que p_1 é menor que p_2.

5.5 - PROBABILIDADE DE ERRO DE CLASSIFICAÇÃO

Vimos anteriormente que a variável

$$U = \underline{X}'\Sigma^{-1}(\underline{\mu}_1-\underline{\mu}_2) - \frac{1}{2}(\underline{\mu}_1+\underline{\mu}_2)'\Sigma^{-1}(\underline{\mu}_1-\underline{\mu}_2)$$

tem distribuição normal, com parâmetros:

$$\underline{X}\in\pi_1 \;\ldots\; E_1(U) = m/2 \;,\; Var_1(U) = m$$

$$\underline{X}\in\pi_2 \;\ldots\; E_2(U) = -m/2 \;,\; Var_2(U) = m$$

onde

$$m = (\underline{\mu}_1-\underline{\mu}_2)'\Sigma^{-1}(\underline{\mu}_1-\underline{\mu}_2)$$

Comentamos também que a distribuição de U é a distribuição limite da estatística V. Admitindo que as presentes amostras são suficientemente grandes podemos considerar V como uma variável aleatória com distribuição normal, cujos parâmetros são calculados utilizando as expressões acima, substituindo-se $\underline{\mu}_i$ e Σ por suas estimativas. Obtemos assim:

$$m = 27,5654$$

$$E_1(V) = 13,7827 \quad , \quad Var_1(V) = 27,5654$$

$$E_2(V) = -13,7827 \quad , \quad Var_2(V) = 27,5654$$

e as probabilidades de erro de classificação podem facilmente ser calculadas com auxílio das tabelas de distribuição normal.

Se $p_1 = p_2 = 0,5$, teremos:

$$P(2|1) = P_1(V<0) = 0,0043$$

$$P(1|2) = P_2(V>0) = 0,0043$$

e a probabilidade de classificar erradamente um elemento escolhido ao acaso será:

$$p_1P(2|1) + p_2P(1|2) = 0,0043.$$

Se $p_1 = 0,0196$ e $p_2 = 0,9804$ teremos:

$$P(2|1) = P_1(V<3,9120) = 0,0301$$

$$P(1|2) = P_2(V>3,9120) = 0,0004$$

e a probabilidade de classificar erradamente um elemento escolhido ao acaso será:

$$p_1P(2|1) + p_2P(1|2) = 0,001.$$

Observamos que no segundo caso a probabilidade de classificar erradamente um indivíduo é bem menor, havendo no entanto um aumento bastante grande de $P(2|1)$. É fácil entender que como a probabilidade de escolher um elemento de π_1 é bem menor que a probabilidade de escolher um elemento de π_2 é compensador o aumento de $P(2|1)$ e a consequente diminuição de $P(1|2)$.

5.6 - UMA REGIÃO DE DÚVIDA

Vamos considerar o caso em que $p_1 = p_2 = 0,5$. Vimos que $P(2|1) = P(1|2) = 0,0043$. Se por algum motivo desejamos limitar estas probabilidades a valores menores que

0,0043 deveremos optar pela partição do espaço amostral proposta por Rao, introduzindo a região de dúvida R_0.

Se por exemplo, a limitação é dada por

$$P(2|1) \leq 0{,}001 \quad e \quad P(1|2) \leq 0{,}001$$

obteremos a seguinte partição

$$R_1 : v \geq A \qquad R_2 : v \leq B \qquad R_0 : B \leq v \leq A$$

onde

$$P(2|1) = P_1(V \leq B) = 0{,}001$$

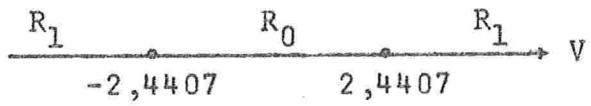
$$P(1|2) = P_2(V \geq A) = 0{,}001$$

Os valores de A e B são obtidos facilmente com auxílio das tabelas de distribuição normal

$$P_1(V \geq B) = 0{,}001 \Longrightarrow B = -2{,}4407$$

$$P_2(V \geq A) = 0{,}001 \Longrightarrow A = 2{,}4407$$

e teremos esquematicamente:



Se um elemento é escolhido ao acaso e o valor de V a ele correspondente varia entre -2,4407 e 2,4407 esse elemento não deve ser classificado em nenhuma das duas populações pois a probabilidade de erro ultrapassa o valor pré-fixado (0,001).

5.7 - A ESCOLHA DAS VARIÁVEIS DISCRIMINANTES

Ao analisar os valores das variáveis Glicose, Insulina e Nefa nas duas amostras (ver ítem 5.2) podemos ter uma primeira idéia do poder discriminante de cada uma dessas variáveis. A primeira a destacar-se à simples observação é a Insulina. Enquanto que na população π_1 os valores da Insulina são altos (de 14 a 22), na população π_2 eles são baixos (de 4 a 14) o que mostra que a Insulina por si só já forneceria uma boa discriminação. A segunda variável a destacar-se é a Nefa que apresenta maiores valores para π_1 do que para π_2. Quanto à Glicose não há uma distinção que salte à vista. Estas conclusões obtidas pela simples análise visual são confirmadas pela comparação dos D^2 de Mahalanobis entre π_1 e π_2 baseados em cada uma das três variáveis.

O D^2 de Mahalanobis com base em uma única variável (X_j) não é nada mais que

$$\frac{(\bar{x}_{1j} - \bar{x}_{2j})^2}{w_{jj}}$$

sendo portanto proporcional à F de uma análise de variância entre π_1 e π_2 com base na variável considerada.

Obteremos para as três variáveis:

Glicose (X_1) ..... $D_1^2 = 0,3796$

Insulina (X_2) ..... $D_2^2 = 19,4577$

Nefa (X_3) ..... $D_3^2 = 4,5787$

o que mostra que a maior distância entre π_1 e π_2 é provocada pela Insulina, seguida em ordem decrescente pela Nefa e Glicose.

Num processo "passo a passo" (stepwise) de seleção de variáveis é evidente que a primeira variável a ser escolhida será a Insulina. (Observe que a discriminação devida à Insulina é significativa). Na escolha da segunda variável, adotaremos o critério já comentado (vide Capítulo 4) de selecionar a variável que introduzir "melhoria de discriminação" mais significativa em relação à discriminação devida às variáveis já selecionadas. Analisemos as melhorias de discriminação introduzidas pela Glicose e pela Nefa.

a) Glicose

O D^2 devido à Glicose e Insulina, $D^2_{1,2}$ é igual a 25,7682. A F correspondente ao teste de melhoria de discriminação é dada

$$F = \frac{(n_1+n_2-p-q-1)n_1n_2(D^2_{1,2}-D^2_1)}{q\{(n_1+n_2)(n_1+n_2-2)+n_1n_2D^2_1\}}$$

onde p é o número de variáveis já consideradas e q o número de variáveis adicionais. No caso p = 1 (Insulina) e q = 1 (Glicose) e o valor de F é 10,0392, significativo ao nível de 5%.

b) Nefa

O D^2 devido à Insulina e à Nefa, $D^2_{2,3}$, é igual a

21,9653. A F correspondente ao teste de melhoria de discriminação, calculado de maneira análoga à descrita para a Glicose, tem o valor 3,9893, não significativo ao nível de 5%, como era de se esperar, pela simples observação dos dados.

Dos resultados acima é imediato que a escolha da segunda variável recai sobre a Glicose. Se ambas as F fossem significativas, consideraríamos a maior delas. Se nenhuma das F fosse significativa, a discriminação se restringiria ao uso da Insulina apenas.

Selecionados a Glicose e Insulina é fácil verificar que a Nefa não introduz melhoria significativa de discriminação. Como a discriminação baseada nas duas primeiras variáveis é significativa podemos nos restringir a elas para fins de discriminação entre π_1 e π_2.

Por exemplo, no caso em que $p_1 = p_2 = 0,5$, as regiões R_1 e R_2 ficam definidas por

$$R_1 : 1,0165x_1 + 2,4212x_2 - 110,8273 \geq 0$$

$$R_2 : 1,0165x_1 + 2,4212x_2 - 110,8273 \leq 0$$

e a probabilidade de classificar erradamente um elemento selecionado ao acaso será 0,0055, o que não representa um aumento sensível comparado com o resultado obtido quando da utilização da Nefa (0,0043).

- □ -

APÊNDICE A

T^2 DE HOTELLING

Em 1931, H. Hotelling publicou um artigo sob o título "Generalization of Student's Ratio" no qual apresentou uma extensão do teste "t" para o caso de populações normais multivariadas. À nova estatística apresentada para tal teste deu a notação de T^2.

De acordo com a notação utilizada neste trabalho temos

$$T^2 = \sum_{i=1}^{p} \sum_{j=1}^{p} w^{ij} \xi_i \xi_j = (n_1+n_2-2) \sum_{i=1}^{p} \sum_{j=1}^{p} S^{ij} \xi_i \xi_j$$

sendo $\xi_i = \lambda d_i$.

À estatística T^2 estão associados n_1+n_2-2 graus de liberdade.

Hotelling demonstrou que a distribuição de

$$\frac{n_1+n_2-p-1}{(n_1+n_2-2)p} T^2$$

sob a hipótese de coincidência das duas populações ($\underline{\mu}_1 = \underline{\mu}_2$) das quais estão sendo extraídas as amostras é simplesmente uma F (de Snedecor) central com p e n_1+n_2-p-1 graus de liberdade.

Um estudo detalhado da estatística T^2 pode ser encontrado em Anderson (referência 1, cap. 5).

D^2 DE MAHALANOBIS

Preocupado com o uso do "coeficiente de semelhança racial" (proposto por K. Pearson em 1921) como métrica para a distância de duas populações, pois este coeficiente era sensivelmente afetado pelo tamanho das amostras, P.C.Mahalanobis começou em 1927 suas pesquisas nessa área, que culminaram com a proposição da estatística D^2, para a estimação do parâmetro populacional Δ^2 (distância generalizada de Mahalanobis).

Note-se que embora o T^2 meça indiretamente distância entre populações, o aspecto que distingue a distância generalizada de Mahalanobis é o fato de ser um parâmetro populacional passível de estimação.

O D^2 é assim definido

$$D^2 = \sum_{i=1}^{p} \sum_{j=1}^{p} w^{ij} d_i d_j = (n_1+n_2-2) \sum_{i=1}^{p} \sum_{j=1}^{p} S^{ij} d_i d_j$$

e

$$\Delta^2 = \sum_{i=1}^{p} \sum_{j=1}^{p} \sigma^{ij} \delta_i \delta_j$$

onde $\delta_i = \mu_{1i} - \mu_{2i}$ e $(\sigma^{ij}) = \Sigma^{-1}$.

É interessante notar que a distância originalmente proposta por Mahalanobis (Δ_M^2) difere desta, incluindo o coeficiente 1/p, onde p é o número de variáveis ($X_1,\ldots,X_p$) consideradas.

$$\Delta_M^2 = \frac{1}{p}\Delta^2$$

A eliminação do fator não afeta em nada o objetivo do uso de Δ^2. A origem desse fator está ligada à influência do coeficiente de semelhança racial de Pearson na origem das pesquisas de Mahalanobis, visto que esse coeficiente também continha o fator.

O D^2 de Mahalanobis se relaciona com o T^2 de Hotelling:

$$T^2 = \sum_{i=1}^{p}\sum_{j=1}^{p} w^{ij}\xi_i\xi_j = \lambda^2\sum_{i=1}^{p}\sum_{j=1}^{p} w^{ij}d_id_j = \lambda^2D^2$$

Disto decorre que a estatística

$$\frac{n_1+n_2-p-1}{(n_1+n_2-2)p}\lambda^2D^2$$

tem distribuição F com p e n_1+n_2-p-1 graus de liberdade.

- □ -

APÊNDICE B

O risco em um processo discriminatório (R) entre k populações é definido por

$$r(R) = \sum_{i=1}^{k} p_i \, r(i,R) = \underline{p}'\underline{r}^R$$

onde

$r(i,R)$: perda esperada (risco) ao classificar um elemento de π_i.

p_i: probabilidade a priori associada a π_i, $(p_1+\ldots+p_k=1)$.

$\underline{p}' = (p_1,\ldots,p_k)$ e $\underline{r}^{R\prime} = (r(1,R),\ldots,r(k,R))$.

Uma regra que minimiza $r(R)$ para um dado vetor $\underline{p}' = (p_1,\ldots,p_k)$ de probabilidades a priori é denominada "regra Bayesiana associada a $\underline{p}$".

Duas regras R e S são *não comparáveis* se para alguns índices i temos $r(i,R) < r(i,S)$ e para alguns j dos índices restantes $r(j,R) > r(j,S)$.

Uma regra R é *melhor* que uma regra S se $r(i,R) \leq r(i,S)$, $i=1,\ldots,k$, e se para pelo menos um i a desigualdade for estrita.

Uma regra R é *admissível* se não houver regra S melhor que R (dentre todas as S comparáveis com R).

Uma classe de regras é *completa* se para toda regra fora da classe, existir uma, dentro da classe, que é melhor.

Uma classe de regras é *completa mínima* se nenhum seu subconjunto próprio for uma classe completa.

Nosso objetivo é mostrar que sob certas condições a classe das regras Bayesianas é completa mínima, o que justifica a escolha de uma regra Bayesiana como solução minimax (ítens 3.3 e 4.3). Para isso, admitiremos sem demonstrar que:

1. A classe dos vetores de perda $\underline{r}^R = (r(1,R),\ldots,r(k,R)),Z$, correspondentes a todas as regras R é um conjunto convexo em $\mathbb{R}^k$.

2. Se C é um conjunto convexo em $\mathbb{R}^k$ tal que os elementos $\underline{x} \in C$ tem pelo menos uma coordenada não negativa, então existe um vetor $\underline{p} \in \mathbb{R}^k$ sem coordenadas negativas e pelo menos uma coordenada positiva tal que $\underline{p}'\underline{x} \geq 0$ para todo $\underline{x} \in C$. (Rao, ref. 10, pag. 42).

3. A classe A de todas as regras admissíveis é completa mínima. (Rao, ref. 10, pag. 418).

<u>TEOREMA 1</u> - Toda regra admissível é uma regra de Bayes com relação a um vetor $\underline{p}$ de probabilidades a priori.

<u>Demonstração:</u>

Seja $\underline{r}^R$ o vetor de perdas associado a uma regra $R \in A$ e $\underline{r}^S$ o vetor de perdas associado a uma regra S. Fixemos R e variemos S no conjunto de todas as regras. O conjunto de ve

tores $\underline{r}^S - \underline{r}^R$ será convexo e nenhum de seus elementos terá todas as componentes negativas pois R é admissível. Logo, por (2) existe um vetor $\underline{p}$ de componentes não negativas e pelo menos uma positiva tal que $\underline{p}'(\underline{r}^S - \underline{r}^R) \geq 0$, isto é, $\underline{p}'\underline{r}^S \geq \underline{p}'\underline{r}^R$, qualquer que seja S e portanto R é uma regra Bayesiana.

TEOREMA 2 - Uma regra Bayesiana com relação a $\underline{p}$ é uma regra admissível se todas as componentes de $\underline{p}$ forem positivas.

Demonstração:

Se R é uma regra de Bayes com relação a $\underline{p}$ teremos

$$\underline{p}'\underline{r}^R \leq \underline{p}'\underline{r}^S, \ \forall S$$

Se existir uma regra A tal que A é melhor que R, teremos

$$\underline{p}'\underline{r}^A \leq \underline{p}'\underline{r}^R \leq \underline{p}'\underline{r}^A$$

e se todas as componentes de $\underline{p}$ são positivas decorre que

$$\underline{r}^R = \underline{r}^A$$

isto é, R é admissível.

Dos teoremas 1 e 2 concluímos que a classe das regras admissíveis e o conjunto das regras Bayesianas associadas a vetores $\underline{p}$ de coordenadas positivas, coincidem e portanto esta regras Bayesianas determinam uma classe completa mínima.

- □ -

APÊNDICE C

TEOREMA - Sejam as funções $F_1 = F_1(x_1,\ldots,x_p),\ldots,F_k = F_k(x_1,\ldots,x_p)$ integráveis em um conjunto $R(R\subset\mathbb{R}^p)$. A partição $R_1,\ldots,R_k$ de R que minimiza a soma

$$\sum_{i=1}^{k}\int_{R_i} F_i d\underline{x} \quad \text{onde} \quad d\underline{x} = dx_1\ldots dx_p$$

é definida por

$$R_i:\ F_i \le F_j \qquad j=1,\ldots,k;\ i\neq j. \tag{1}$$

Demonstração:

Sejam $R_1',\ldots,R_k'$: uma partição qualquer de R

e $R_1,\ldots,R_k$: uma partição de R definida por (1)

Se $R_{ij} = R_i \cap R_j'$ teremos $\bigcup_{j=1}^{k} R_{ij} = R_i$ e $\bigcup_{i=1}^{k} R_{ij} = R_j'$.

Da definição de R_i decorre que:

$$\int_{R_i} F_i d\underline{x} \le \int_{R_{i1}} F_1 d\underline{x} + \ldots + \int_{R_{ik}} F_k d\underline{x}$$

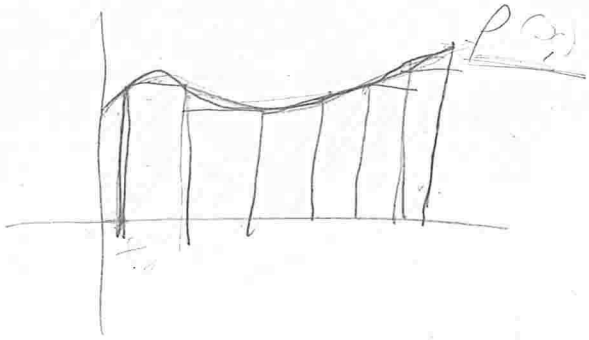
Somando para $i=1,\ldots,k$, teremos

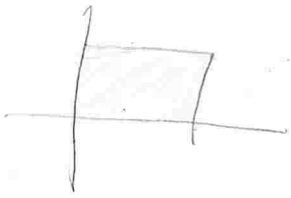
$$\sum_{i=1}^{k}\int_{R_i} F_i d\underline{x} \le \sum_{i=1}^{k}\int_{R_{i1}} F_1 d\underline{x} + \ldots + \sum_{i=1}^{k}\int_{R_{ik}} F_k d\underline{x}$$

e portanto

$$\sum_{i=1}^{k} \int_{R_i} F_i d\underline{x} \leq \sum_{i=1}^{k} \int_{R_i'} F_i d\underline{x}.$$

- □ -





APÊNDICE D

LEMA DE NEYMAN-PEARSON - Sejam as funções $F_0 = F_0(x_1,\ldots,x_p)$, $\ldots, F_k = F_k(x_1,\ldots,x_p)$ funções integráveis em $\mathbb{R}^p$.
Seja $C(C \subset \mathbb{R}^p)$ uma região para a qual

$$\int_C F_i d\underline{x} = \alpha_i \quad i=1,\ldots,k \tag{1}$$

onde $d\underline{x} = dx_1 \ldots dx_p$ e α_i é uma constante pré-determinada.

De todas as regiões para as quais (1) é válida, a que maximiza

$$\int_C F_0 d\underline{x}$$

é definida por C: $F_0 \geq a_1 F_1 + \ldots + a_k F_k$ onde as constantes $a_1,\ldots,a_k$ são tais que a restrição (1) é satisfeita.

Demonstração:

Seja C' uma região na qual (1) é válida e C uma região tal que são válidas (1) e $F_0 \geq a_1 F_1 + \ldots + a_k F_k$. Seja CC' a intersecção de C e C'.

É imediato que

$$\int_{C-CC'} F_i d\underline{x} = \int_{C'-CC'} F_i d\underline{x} \quad i=1,\ldots,k \tag{2}$$

Consideremos a diferença D:

$$D = \int_C F_0 d\underline{x} - \int_{C'} F_0 d\underline{x} = \int_{C-CC'} F_0 d\underline{x} - \int_{C'-CC'} F_0 d\underline{x} \geq \int_{C-CC'} (\sum a_i F_i) d\underline{x} - \int_{C'-CC'} (\sum a_i F_i) d\underline{x}$$

de (2) resulta que

$$D \geq 0$$

ou

$$\int_C F_0 d\underline{x} \geq \int_{C'} F_0 d\underline{x}.$$

- □ -

BIBLIOGRAFIA

[1] - ANDERSON, T.W. - (1958) - An Introduction to Multivariate Statistical Analysis - Wiley.

[2] - BENETT, C.A. e FRANKLIN, N.L. - (1967) - Statistical Analysis in Chemistry and The Chemical Industry - Wiley.

[3] - FISHER, R.A. - (1936) - The Use of Multiple Measurments in Taxonomic Problems - Annals of Eugenics, Vol. VII, Pt. II.

[4] - FISHER, R.A. - (1938) - The Statistical Utilization of Multiple Measurments - Annals of Eugenics, Vol. VIII, Parte IV.

[5] - GOULDEN, C.H. - (1952) - Methods of Statistical Analysis - Wiley.

[6] - HODGES, J.L. - (1955) - Discriminatory Analysis-USAF School of Aviation Medicine, Texas.

[7] - KENDALL, M.G. e STUART, A. - (1966) - The Advanced Theory of Statistics - Vol. 3 - Griffin.

[8] - OVERALL, J.E. e KLETT, C.I. - (1972) - Applied Multivariate Analysis - McGraw Hill.

[9] - RAO, C.R. - (1952) - Advanced Statistical Methods in Biometric Research - Wiley.

[10] - RAO, C.R. - (1965) - Linear Statistical Inference and its Applications - Wiley.

[11] - RAO, C.R. - (1972) - Recent Trends of Research Work in Multivariate Analysis - Biometrics 28, 3-22, March.

[12] - ROMEDER, J.M. (1973) - Méthodes et Programmes d'Analyse Discriminante - Dunod.

[13] - STEVENS, W.L. - (1945) - Questões de Método - Universidade de Coimbra.

[14] - TATSUOKA, M.M. e TIEDEMAN, D.V. - (1954) - Discriminant Analysis - Review of Educational Research - Vol. XXIV, 5.

OBSERVAÇÃO: Os artigos 3 e 4 podem ser encontrados no livro:

FISHER, R.A. - (1950) - Contributions to Mathematical Statistics - Wiley.

- □ -