Luiz Paulo Fávero
Patrícia Belfiore
Fabiana Lopes da Silva
Betty Lilian Chan

análise de dados

MODELAGEM MULTIVARIADA PARA TOMADA DE DECISÕES

CAMPUS

CAPÍTULO 11

Análise Discriminante

Para saber que nós sabemos o que nós sabemos,
e para saber que nós não sabemos o que nós não sabemos,
isto é o conhecimento verdadeiro.

NICOLAU COPÉRNICO

AO FINAL DESTE CAPÍTULO, VOCÊ SERÁ CAPAZ DE:

- Prescrever as situações em que uma função discriminante deve ser utilizada.
- Saber diferenciar as circunstâncias nas quais a análise discriminante deve ser usada, em vez da regressão múltipla.
- Definir questões apropriadas de pesquisá abordadas pela análise discriminante, compreendendo as suposições relativas à técnica.
- Identificar questões importantes para a aplicação da análise discriminante.
- Entender as abordagens computacionais para a análise discriminante e saber diferenciar o método simultâneo do método *stepwise*.
- Compreender a natureza das funções discriminantes.
- Identificar as variáveis que discriminam os grupos previamente definidos, utilizando o poder discriminatório significante.
- Elucidar os resultados da análise discriminante por meio de representações gráficas.
- Reconhecer o perfil de cada grupo apresentado.

11.1. APRESENTAÇÃO DO CAPÍTULO[1]

A análise discriminante (AD) é uma técnica multivariada utilizada quando a variável dependente é categórica, ou seja, qualitativa (não métrica) e as variáveis independentes são quantitativas (métricas).

Como objetivo principal, a análise discriminante oferece ao pesquisador a possibilidade de elaborar previsões a respeito de a qual grupo certa observação (por exemplo, um produto, uma pessoa ou uma empresa) pertencerá, uma vez que se caracteriza como uma técnica de previsão e classificação. Para alcançar este objetivo, a análise discriminante gera funções discriminantes (combinações lineares das variáveis) que ampliam a discriminação dos grupos descritos pelas categorias de determinada variável dependente.

[1] Agradecemos a Débora Confortini por sua importante contribuição quando da elaboração deste capítulo.

Definida esta proposição, um pesquisador pode estar interessado no esclarecimento das correlações que motivam as características de cada categoria na qual o objeto está posicionado, tal como o estudo de uma companhia de seguros que apresenta o intuito de prever quais clientes estão mais predispostos à falência. Outro pesquisador pode desejar avaliar a percepção do departamento de vendas de uma empresa sobre o fato de um novo produto estar destinado ao sucesso ou ao fracasso quando do seu lançamento. Um terceiro pode ainda investigar as razões por meio das quais há a liberação de crédito por uma empresa do setor financeiro e em definir critérios para a estratificação de pessoas em grupos com características de bons pagadores com acesso total ao crédito, bons pagadores com restrição ao crédito e maus pagadores sem direito a crédito.

Em resumo, a análise discriminante ajuda o pesquisador que deseja criar um estudo com grupos diferenciados por meio das variáveis independentes que possui. Em outras palavras, a análise discriminante comporta-se como uma técnica confirmatória da análise de conglomerados estudada no Capítulo 6.

Os três maiores objetivos deste capítulo são: (1) introduzir a natureza, a filosofia e as condições de uso da análise discriminante; (2) expressar a aplicabilidade da técnica; e (3) discutir os resultados obtidos por meio da utilização de *softwares* estatísticos.

11.2. UMA INTRODUÇÃO À ANÁLISE DISCRIMINANTE

A análise discriminante foi proposta na primeira metade do século XX por Fisher, para servir como um critério mais confiável para a classificação de novas espécies de vegetais de acordo com as características biométricas, sendo rapidamente adotada além da Taxonomia e Sistemática Vegetal (MAROCO, 2007). Percebe-se uma grande aplicação da AD em diversos campos do conhecimento, como biologia, antropologia, marketing, comportamento do consumidor, entre outros. Em finanças, especificamente, merece destaque o seminal trabalho de Edward Altman que, em 1968, publicou no **The Journal of Finance** o artigo denominado *Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy*. Neste trabalho, Altman ainda cita outros autores que também contribuíram significativamente para a aplicação da AD.

Durante muitos anos, a análise discriminante tem recebido uma grande atenção teórica de diversas áreas, como marketing, em que podem ser citados os trabalhos de Frank, Massy e Morrison (1965), Morrison (1969), Crask e Perreault (1977) e Hora e Wilcox (1982). Merece destaque também a expressiva contribuição de alguns trabalhos em relação à modelagem matemática da análise discriminante, como os de Lachenbruch e Mickey (1968), Marks e Dunn (1974), McLachlan (1974), Krzanowski (1975), Randles, Broffitt, Ramberg e Hogg (1978), Constanza e Afifi (1979) e Fraley e Raftery (2002).

Nos Estados Unidos e na Europa, há, atualmente, uma vasta aplicabilidade da análise discriminante nas ciências sociais e do comportamento e, no Brasil, seu uso vem sendo ampliado em diversas áreas, devido à contribuição direta dos principais *softwares* estatísticos que apresentam esta técnica.

A análise discriminante envolve a relação entre o conjunto de variáveis independentes quantitativas e uma variável dependente qualitativa. Em muitos casos, verificam-se mais de três classificações para a variável dependente (neste caso, multicotômica), como, por exemplo, classificações do tipo alto, médio e baixo ou bom pagador sem restrição a crédito, bom pagador com restrição a crédito e mau pagador com total restrição a crédito. Neste capítulo, vamos exemplificar a análise discriminante por meio de um banco de dados cuja variável dependente apresenta três categorias (multicotômica).

Quando o pesquisador estiver interessado na discussão de somente dois grupos de variáveis dependentes, a técnica é chamada de Análise Discriminante Simples. No entanto, em muitos casos, há o interesse na discriminação entre mais de dois grupos, sendo a técnica, assim, denominada de Análise Discriminante Múltipla (a partir de agora, chamaremos de MDA – *Multiple-group discriminant analysis*).

Os objetivos principais desses dois tipos de análise são parecidos: (i) identificar as variáveis que melhor discriminam dois ou mais grupos; (ii) utilizar estas variáveis para desenvolver funções discriminantes que representam as diferenças entre os grupos; (iii) fazer uso das funções discriminantes para o desenvolvimento de regras de classificação de futuras observações nos grupos.

A Análise Discriminante Simples requer somente uma função discriminante para representar todas as diferenças entre os dois grupos. Por outro lado, a MDA oferece, como *output*, funções adicionais que conseguem explicitar as diferenças entre os grupos. Sendo assim, a MDA também tem, como objetivo adicional: (iv) identificar o número mínimo de funções discriminantes que melhor proporciona as diferenças entre os grupos (SHARMA, 1996). Merece atenção o fato de a MDA demandar uma quantidade de variáveis independentes maior ou igual ao número de grupos (categorias) da variável dependente em estudo. O pesquisador ainda perceberá que a quantidade de funções discriminantes criadas para a MDA representa o número de grupos (categorias da variável dependente) menos um.

$$\text{Nº. de funções discriminantes} = (g - 1) \tag{11.1}$$

Em que g representa a quantidade de categorias (grupos) da variável dependente.

A Figura 11.1 apresenta os tipos de análise discriminante em função da quantidade de categorias da variável dependente.

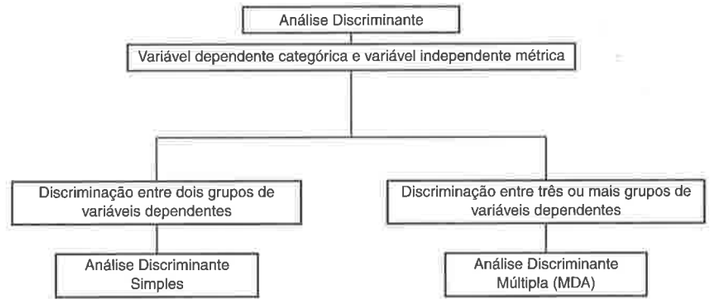


Figura 11.1: Tipos de análise discriminante.

11.3. MODELAGEM DA ANÁLISE DISCRIMINANTE

Antes de iniciarmos a modelagem da análise discriminante propriamente dita, é pertinente esclarecermos os pressupostos inerentes a esta técnica. Os *softwares* estatísticos atualmente apresentam os cálculos dos testes referentes a esses pressupostos e, portanto, é interessante que o pesquisador avalie os resultados e compare com os níveis padrão de utilização, para avaliar se a análise discriminante em questão caracteriza-se por estar em um nível confiável de aplicação, ou seja, se não fere alguns dos critérios estabelecidos por meio de suas premissas. No SPSS, esses testes aparecem antes da projeção da função discriminante, permitindo, assim, essa comparação.

Pode-se dizer que há dois pressupostos principais, referentes à existência de normalidade multivariada das variáveis explicativas e à presença de homogeneidade das matrizes de variância e covariância para os grupos. Em relação à primeira suposição, a combinação linear das variáveis explicativas apresenta uma distribuição normal e, caso ocorra uma violação desse pressuposto, a AD poderá causar distorções nas avaliações do pesquisador, principalmente se a amostra que compõe cada grupo for pequena.

No entanto, se essa violação somente ocorrer pela existência de assimetria da distribuição, a decisão sobre a aplicação da técnica não sofrerá alteração. É importante ressaltar que, se a distribuição não for mesocúrtica, a aplicação da AD será prejudicada, sendo pior o caso em que a distribuição multivariada for

platicúrtica (SHARMA, 1996). Na página virtual do livro na Web (www.campus.com.br) está apresentada a rotina para aplicação do teste de normalidade multivariada por meio do pacote computacional SAS.

O segundo pressuposto refere-se à existência de homogeneidade das matrizes de variância e covariância. Esse pressuposto é verificado por meio da estatística Box's M, que pode ser sensível ao tamanho da amostra e ao não-atendimento da hipótese de distribuição normal multivariada. Felizmente, a análise discriminante é uma técnica bastante robusta à violação desses pressupostos, desde que a dimensão do menor grupo seja superior ao número de variáveis em estudo e que as médias dos grupos não sejam proporcionais às suas variâncias, ou seja, caso a homogeneidade das matrizes de variância e covariância seja violada, haverá um aumento da probabilidade para classificar observações no grupo que possuir a maior dispersão.

Além desses pressupostos, é pertinente ressaltar que a inexistência de *outliers*, a presença de linearidade das relações e a ausência de problemas relacionados à multicolinearidade das variáveis explicativas também são consideradas pressupostos da análise discriminante.

É essencial definirmos o tamanho correto da amostra que será estudada, uma vez que esta técnica é muito sensível à proporção do tamanho da amostra em relação ao número de variáveis preditoras (HAIR, ANDERSON, TATHAM e BLACK, 2005) e, portanto, não deve haver uma grande variabilidade de dimensões entre os grupos. Como regra geral, utilizam-se 20 observações para cada variável explicativa, mesmo que o número final de variáveis preditoras a serem incluídas no modelo seja reduzido (método *stepwise*).

Ainda segundo Hair, Anderson, Tatham e Black (2005), como a dimensão do menor grupo deve exceder o número de variáveis explicativas, muitos recomendam que cada grupo também tenha um mínimo de 20 observações. Porém, ressaltamos que a definição do dimensionamento amostral seja estabelecida de acordo com algum critério e sempre embasada pelos conceitos de amostragem apresentados no Capítulo 5.

Apresentados os pressupostos, podemos começar a expor os passos para a composição das funções discriminantes, lembrando que isso representa um dos principais objetivos da AD, já que, a partir destas funções é que as observações serão discriminadas. Antes desta primeira etapa, lembramos que o pesquisador já deve ter estabelecido seu problema principal de pesquisa, assim como em qualquer aplicação de uma técnica multivariada, e definido as categorias de estudo com, no mínimo, dois grupos.

Tomemos, como exemplo, uma loja que deseja elaborar um estudo sobre a concessão do limite de crédito aos clientes (problema principal) quando da elaboração do crediário e, para tanto, deverá discriminá-los em dois grupos, sendo, respectivamente, o dos inadimplentes e o dos adimplentes (níveis de estudo divididos). Neste caso, estamos tratando de uma AD em nível dicotômico.

Portanto, esta etapa consiste na seleção da variável dependente (categórica) e das variáveis explicativas (métricas). Em outras palavras, a escolha de n variáveis discriminantes (explicativas) é feita a partir de um conjunto maior de p possíveis variáveis. Todas as etapas e os testes aqui discutidos serão apresentados e detalhados quando da elaboração da modelagem da AD por meio do pacote computacional SPSS, nas Seções 11.6 e 11.7.

A análise discriminante permite o conhecimento das variáveis que mais se destacam na discriminação dos grupos. Para tanto, diversos *outputs* são gerados a partir de testes e estatísticas, como o lambda de Wilks, a correlação canônica, o Qui-quadrado e o *eigenvalue*.

O lambda de Wilks, que varia de 0 a 1, propicia a avaliação da existência de diferenças de médias entre os grupos para cada variável. Valores elevados desta estatística indicam ausência de diferenças entre os grupos, e sua expressão é dada por:

$$\Lambda = \frac{SQ_{dg}}{SQT} \tag{11.2}$$

em que SQ_{dg} representa a soma dos erros (dentro dos grupos) e SQT, a soma dos quadrados total.

Como a distribuição exata de lambda não é conhecida, utiliza-se, para a existência de dois grupos, a seguinte transformação, que possui distribuição F com p e $(N-p-1)$ graus de liberdade, em que N representa o tamanho da amostra e p o número total de variáveis explicativas.

$$F = \left(\frac{1-\Lambda}{\Lambda}\right)\left(\frac{N-p-1}{p}\right) \tag{11.3}$$

ou a seguinte transformação, para a existência de três grupos, com distribuição F com $2p$ e $2(N-p-2)$ graus de liberdade (MAROCO, 2007):

$$F = \left(\frac{1-\sqrt{\Lambda}}{\sqrt{\Lambda}}\right)\left(\frac{N-p-2}{p}\right) \tag{11.4}$$

O valor transformado de lambda de Wilks segue uma exata distribuição F somente em certos casos, apresentados no Quadro 11.1. Para todos os outros casos, a distribuição do valor transformado de lambda de Wilks pode somente ser aproximado por uma distribuição F (SHARMA, 1996).

Quadro 11.1: Situações em que o Lambda de Wilks apresenta Distribuição F

Número de Variáveis (p)	Número de Grupos	Transformação	Graus de Liberdade
n	2	$\left(\frac{1-\Lambda}{\Lambda}\right)\left(\frac{N-p-1}{p}\right)$	$p, N-p-1$
n	3	$\left(\frac{1-\sqrt{\Lambda}}{\sqrt{\Lambda}}\right)\left(\frac{N-p-2}{p}\right)$	$2p, 2(N-p-1)$
1	qualquer	$\left(\frac{1-\Lambda}{\Lambda}\right)\left(\frac{N-g}{g-1}\right)$	$g-1, N-g$
2	qualquer	$\left(\frac{1-\sqrt{\Lambda}}{\sqrt{\Lambda}}\right)\left(\frac{N-g-1}{g-1}\right)$	$2(g-1), 2(N-g-1)$

Fonte: Sharma (1996).
Nota: g = número de grupos.

Com a seleção das variáveis discriminantes (explicativas) para a formação dos grupos, passamos à identificação das funções discriminantes. Como já brevemente discutido, a AD assemelha-se à análise de regressão em termos de objetivos e características e, desta forma, sua função geral pode ser representada por meio da seguinte equação linear:

$$Z_n = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n \tag{11.5}$$

em que:
Z: variável dependente;
α: intercepto;
X_i: variáveis explicativas;
β_i: coeficientes discriminantes para cada variável explicativa.

É importante ressaltar que esta função discriminante é diferente da função discriminante linear de Fisher, uma vez que, enquanto a primeira é utilizada como um meio de facilitar a interpretação dos parâmetros das variáveis explicativas, a função discriminante linear de Fisher é utilizada para classificar as observações nos grupos.

Na função discriminante linear de Fisher, os valores das variáveis explicativas de uma observação são inseridos nas funções de classificação e, consequentemente, um escore de classificação é calculado para cada grupo, para aquela observação.

Dadas p variáveis e g grupos, é possível estabelecer $m = \min(g\text{-}1; p)$ funções discriminantes que são combinações lineares das p variáveis, de modo que a função linear de Fisher seja dada por:

$$Z_n = W_1X_1 + W_2X_2 + \ldots + W_nX_n \tag{11.6}$$

em que W_i representa o vetor de pesos das variáveis para as funções discriminantes e são estimados de modo que a variabilidade dos escores da função discriminante seja máxima entre os grupos e mínima dentro dos grupos (MAROCO, 2007). Assim, podemos expressar o i-ésimo *eigenvalue* (i-ésima função discriminante) da seguinte forma:

$$eigenvalue_i = \frac{SQ_{eg}}{SQ_{dg}} \tag{11.7}$$

em que:
SQ_{eg}: soma dos quadrados entre os grupos;
SQ_{dg}: soma dos quadrados dentro dos grupos.

Assim, *eigenvalues* altos resultam em boas funções discriminantes.

Expressa de acordo com a Equação (11.6), a função discriminante é conhecida por função discriminante linear de Fisher e, após a dedução da primeira função discriminante, os pesos W_i das funções seguintes são obtidos sob a restrição adicional de que os escores das funções discriminantes não estejam correlacionados.

Por meio desses cálculos, é possível chegarmos à expressão da primeira função discriminante (Z_1); as outras funções dos outros grupos são encontradas pelo mesmo método. Porém, é preciso que tenhamos atenção em relação a problemas de correlação, ou seja, os escores das outras funções (Z_2, Z_3, ..., Z_n) não devem ser correlacionados. Assim, a expressão do *eigenvalue* da segunda função discriminante linear de Fisher pode ser escrito da seguinte forma:

$$eigenvalue = \frac{SQ_{eg}(Z_2)}{SQ_{dg}(Z_2)} \tag{11.8}$$

em que:

$SQ_{eg}(Z_2)$: soma dos quadrados entre os grupos na segunda função discriminante;
$SQ_{dg}(Z_2)$: soma dos quadrados dentro dos grupos na segunda função discriminante.

Por meio das funções discriminantes lineares de Fisher, é possível estudarmos a influência que determinada variável dependente categórica sofre de um vetor de variáveis explicativas. Ademais, é possível estabelecer uma relação entre os *eigenvalues* e o lambda de Wilks:

$$\Lambda = \Pi\left[\frac{1}{(1+eigenvalue_i)}\right] \tag{11.9}$$

Como já discutido anteriormente, para a existência de g grupos, há (g-1) funções discriminantes. Assim, se tivéssemos, por exemplo, três grupos, seriam geradas duas funções discriminantes, sendo a primeira responsável pela separação de um grupo dos demais e a segunda pela separação dos dois grupos restantes.

Como nem todas as funções discriminantes podem ser estatisticamente significantes, ou seja, algumas podem representar maiores diferenças entre os grupos do que outras, é necessária a seguinte transformação que possui distribuição Qui-quadrado (χ^2) com $p.(g\text{-}1)$ graus de liberdade. A expressão a seguir é utilizada, portanto, para permitir o cálculo da estatística Qui-quadrado (χ^2) que avalia a significância estatística global de todas as funções discriminantes (SHARMA, 1996):

$$\chi^2 = -\left[n - \frac{(p+g)}{2} - 1\right].\ln(\Lambda_k) \tag{11.10}$$

em que Λ_k é o lambda de Wilks de cada função discriminante e testa a significância das funções discriminantes, ou seja, avalia o quão bem cada função separa as observações em grupos diferentes. Se tivermos, por exemplo, duas funções discriminantes, primeiramente são testadas as duas funções em conjunto e, na sequência, se as médias dos grupos para a segunda função discriminante são iguais. Nesta etapa, a rejeição da hipótese nula do teste Qui-quadrado (H_0: nenhuma das funções é significante para discriminar os grupos) não significa que tenhamos condições de responder quais (ou qual) funções discriminam significativamente os grupos, ou seja, a rejeição de H_0 significa apenas que a primeira função discriminante é significante (as outras podem não ser).

Outra estatística resultante refere-se à correlação canônica, que corresponde à razão entre a variação entre os grupos e a variação total e mede o grau de associação entre os escores discriminantes e os grupos. Sua expressão é dada por:

$$\text{CANCOR} = \sqrt{\frac{SQ_{eg}}{SQT}} \tag{11.11}$$

que resulta em:

$$\Lambda + CANCOR^2 = 1 \tag{11.12}$$

Na Seção 11.6 será elaborada uma aplicação da AD e, a partir daí, tais expressões serão utilizadas e os respectivos resultados confrontados.

Em termos matriciais, Sharma (1996) apresenta a função discriminante como:

$$\xi = X'\gamma \tag{11.13}$$

em que X' ($p \times 1$) é a transposta da matriz com p variáveis e γ representa o vetor de pesos das variáveis. A soma dos quadrados totais para os escores ξ pode ser definido como $\xi'\xi = (X'\gamma)'(X'\gamma) = \gamma' XX'\gamma$, sendo $XX'=T$ a matriz da soma de quadrados e produtos cruzados totais da matriz X com p variáveis.

Fazendo $T = B + W$, em que B e W representam, respectivamente, as matrizes das somas dos quadrados entre os grupos e dentro dos grupos, a soma dos quadrados totais para a função discriminante pode agora ser escrita como $\xi'\xi = \gamma' T\gamma = \gamma'(B + W)\gamma = \gamma' B\gamma + \gamma' W\gamma$ (MAROCO, 2007).

Uma vez que $\gamma' B\gamma$ e $\gamma' B\gamma$ são, respectivamente, a soma dos quadrados entre os grupos e dentro dos grupos para a função ξ, a obtenção da função discriminante resume-se, segundo Maroco (2007), a encontrar o vetor γ, de modo que:

$$\lambda = \frac{\gamma' B\gamma}{\gamma' W\gamma} \tag{11.14}$$

seja máximo.

Sharma (1996) explica que o problema dessa maximização apresenta solução quando:

$$(W^{-1}B - \lambda I) = 0 \tag{11.15}$$

sob a restrição $|W^{-1}B - \lambda I| = 0$.

Este problema tem m = min (k-1; p) soluções correspondentes aos *eigenvalues* da matriz $W^{-1}B$. O maior *eigenvalue* (λ_1) apresenta um vetor próprio (*eigenvector*) correspondente à primeira função discriminante, o segunda maior *eigenvalue* (λ_2) apresenta um vetor próprio correspondente à segunda função, e assim sucessivamente, sob a restrição de que os escores das funções não estejam correlacionados (STEVENS, 2002).

Apresentamos agora as hipóteses nula e alternativa que se referem às médias populacionais dos grupos em análise para as variáveis explicativas. Para tanto, utilizamos a estatística da distribuição F para testar as seguintes hipóteses:

$$H_0: \mu_1 = \mu_2 = \ldots = \mu_n$$
$$H_1: \mu_1 \neq \mu_2 \neq \ldots \neq \mu_n$$

em que μ_1, μ_2, ..., μ_n são as médias populacionais dos grupos 1, 2, ..., n, respectivamente. A hipótese nula será rejeitada se pelo menos as médias de dois grupos forem significativamente diferentes (SHARMA, 1996). Desta forma, a hipótese alternativa, quando não rejeitada, indica que as variáveis explicativas apresentam médias diferentes entre os grupos.

Após a função discriminante ser definida, será calculado o escore discriminante da variável dependente (Z) para cada observação, ou seja, os escores serão calculados de maneira a propiciar a definição do escore crítico que determinará a forma por meio da qual iremos classificar uma observação em determinado grupo. Para grupos de mesma dimensão amostral (tamanho), o cálculo do escore de corte (*cutoff value*) é:

$$f = \frac{\bar{d}_1 + \bar{d}_2}{2} \tag{11.16}$$

em que $\bar{d}_1$ e $\bar{d}_2$ representam as médias das funções discriminantes (centróides) nos grupos 1 e 2, respectivamente. Já para grupos com tamanhos diferentes, temos:

$$f = \frac{n_1\bar{d}_1 + n_2\bar{d}_2}{n_1 + n_2} \tag{11.17}$$

em que n_1 e n_2 são os tamanhos dos grupos 1 e 2, respectivamente. Normalmente, o valor de corte selecionado é aquele que minimiza o número de classificações incorretas (SHARMA, 1996). De acordo com Maroco (2007), um caso é classificado no grupo 1 se seu escore na função discriminante for maior que f. No caso de mais de dois grupos, pode-se definir a zona de fronteira para cada par de grupos, conhecida por mapa territorial. Como os centróides representam as médias das funções discriminantes em cada grupo, temos, em determinado mapa territorial, que:

2 grupos = 2 centróides
3 grupos = 3 centróides
4 grupos = 4 centróides
...

O método do cálculo do escore crítico é um dos procedimentos existentes para a classificação de futuras observações. Outros métodos também podem ser mencionados, como o da Teoria da Decisão Estatística, o da Função de Classificação e o D^2 de Mahalanobis.

Maroco (2007) descreve que a distância de cada escore ao centróide de um grupo pode ser calculado da seguinte forma:

$$D_j^2 = (d - \bar{d}_j)S_j^{-1}(d - \bar{d}_j)^t \tag{11.18}$$

em que S_j^{-1} representa a matriz de variância e covariância para as funções discriminantes no grupo g e $\bar{d}_j$ o centróide deste grupo. Se uma observação pertence ao grupo j (j = 1, ..., g), então D_g^2 possui distribuição Qui-quadrado com m (número de funções discriminantes) graus de liberdade.

11.4. PROCEDIMENTOS COMPUTACIONAIS PARA ELABORAÇÃO DA ANÁLISE DISCRIMINANTE: SIMULTÂNEO E *STEPWISE*

Nos *softwares* estatísticos, como o SPSS, existem dois procedimentos para a definição das funções discriminantes: simultâneo e *stepwise*.

Como o próprio nome diz, o procedimento simultâneo considera a inclusão de todas as variáveis explicativas conjuntamente no modelo, mesmo quando uma ou mais delas não forem significativas, assim como já estudado na técnica de regressão múltipla no Capítulo 10.

O procedimento *stepwise*, por outro lado, é utilizado quando o pesquisador deseja avaliar a significância estatística das variáveis por meio da inclusão passo a passo apenas das variáveis significantes. É como se a análise começasse sem variável explicativa alguma e, conforme os parâmetros das variáveis vão sendo testados, elas podem ser adicionadas ou não ao modelo. No início do procedimento, ocorre a escolha da melhor variável explicativa discriminante dos grupos e, na sequência, outras variáveis vão sendo testadas a partir da significância de seus parâmetros. O procedimento *stepwise*, que só termina quando não há mais nenhuma variável a ser adicionada ou excluída, é bastante útil quando o pesquisador quer considerar um número relativamente grande de variáveis explicativas para inclusão na função. Este procedimento oferece diversos métodos de seleção (inclusão ou exclusão) de variáveis discriminantes na função discriminante e, entre eles, merecem destaque o método de lambda de Wilks, o D^2 de Mahalanobis, o Smallest F ratio (Razão F entre Grupos), o V de Rao e o método Unexplained Variance.

Lambda de Wilks: esta estatística propicia a inclusão ou exclusão de variáveis de acordo com seu valor lambda. Sua expressão é dada por:

$$\Lambda = \frac{SQ_{dg}}{SQT} \tag{11.19}$$

em que SQ_{dg} representa a soma dos erros (dentro dos grupos) e SQT, a soma dos quadrados total, exatamente como já apresentado por meio da Expressão (11.2).

Portanto, as variáveis a serem incluídas maximizam significativamente as diferenças entre os grupos e minimizam a heterogeneidade dentro dos grupos (MAROCO, 2007). É importante lembrar que pode ser elaborada uma aproximação com a distribuição F, de acordo com o apresentado por meio do Quadro 11.1.

D^2 de Mahalanobis: quando existem mais de dois grupos e as variáveis apresentam correlações significativas entre si, a inclusão da covariância ou correlação entre as variáveis na análise pode conduzir a uma maior separação entre os grupos. A distância de Mahalanobis entre duas variáveis i e j é (MAROCO, 2007):

$$D = \sqrt{(X_i - X_j)^t S^{-1}(X_i - X_j)} \tag{11.20}$$

A distância D^2 de Mahalanobis é utilizada de modo a assegurar ainda mais a separação de todos os grupos.

Razão F entre Grupos: é a conversão da distância de Mahalanobis em uma razão F entre os grupos. Segundo Maroco (2007), sua expressão, para dois grupos *a* e *b*, é dada, por:

$$F_{ab} = \frac{(n-p-g)n_a n_b}{p(n-p)(n_a - n_b)} D^2 \qquad (11.21)$$

Esta transformação da distância de Mahalanobis considera as diferentes dimensões entre os grupos, de modo que os grupos de maior dimensão apresentem maior peso na análise.

V de Rao: como descreve Sharma (1996), esta estatística é baseada na distância de Mahalanobis e busca maximizar a distância entre os centróides dos grupos e o centróide geral. Sua expressão é dada por:

$$V = -(n-g)\sum_{i=1}^{p}\sum_{j=1}^{p} W_{ij}^{-1}(t_{ij} - W_{ij}) \qquad (11.22)$$

em que *n* é a dimensão da amostra, *g* é o número de grupos e w_{ij} e t_{ij} representam os elementos genéricos da matriz W e T, respectivamente. Podemos elaborar sua aproximação com a distribuição Qui-quadrado com (*g*-1) graus de liberdade, a fim de descobrirmos a significância estatística para inclusão ou exclusão de determinada variável. O uso do V de Rao, ao contrário do lambda de Wilks, não procura maximizar a homogeneidade dentro dos grupos, considerando somente a maximização da heterogeneidade entre os grupos (MAROCO, 2007).

Unexplained Variance: adiciona somente as variáveis que minimizam a soma da variação não explicada entre os grupos.

O pesquisador deve ter em mente que nenhum método de seleção de variáveis é preferível aos demais e que estes chegam, por vezes, a soluções diferentes. O método de lambda de Wilks tem sido o mais comumente encontrado em pesquisas recentes.

11.5. ANÁLISE DISCRIMINANTE E PROCEDIMENTO SIMULTÂNEO: UM EXEMPLO PRÁTICO

O procedimento simultâneo só difere do *stepwise* em relação ao método de inclusão ou exclusão de variáveis explicativas. Porém, para a elaboração de problemas de pesquisa (com variável dependente categórica e variáveis explicativas métricas) e escolha final das variáveis a serem incluídas no modelo, os dois procedimentos são iguais.

Como exemplo, consideremos um estudo feito por uma empresa interessada em pesquisar sobre a satisfação no emprego de seus funcionários (problema de pesquisa). Cada entrevistado deu uma nota de 0 a 20 pelo trabalho a que é submetido e, então, foi classificado (categorizado) de acordo com esta nota, sendo estabelecido o seguinte critério:

- muito satisfeito se desse nota superior a 14;
- moderadamente satisfeito se a nota fosse de 7 a 14;
- pouco satisfeito se a nota fosse menor do que 7.

Este exemplo é destinado, exclusivamente, a fins didáticos, sendo os dados puramente ilustrativos.

No mesmo questionário, outras variáveis foram pesquisadas, como o salário atual, o salário inicial na organização, o tempo de serviço (expresso em meses desde o primeiro dia de trabalho na empresa), a escolaridade (anos de estudo) e a idade do entrevistado. Estas representam variáveis explicativas que poderão formar a função discriminante, a fim de que os casos sejam distribuídos nos grupos corretos.

Como já comentado, faremos uso do *software* estatístico SPSS para a elaboração da AD.

11.5.1. Preparação da Modelagem

A partir das respostas apresentadas pelos entrevistados, foi estruturado um banco de dados no SPSS. A descrição de cada variável é apresentada no Quadro 11.2 a seguir:

Quadro 11.2: Variáveis Explicativas e Respectivos *Labels*

Variável	*Label*
nota	Nota emitida pelo Empregado
satisfa_emg	Classificação de Satisfação
sal_atual	Salário Atual ($)
sal_inicial	Salário Inicial na Organização ($)
tempo_ser	Tempo de Serviço (meses)
nivel_educ	Anos de Estudo
idade	Idade do Pesquisado (anos)

Foram entrevistados todos os 32 funcionários da empresa. Lembramos que esta dimensão amostral é pequena em relação aos critérios apresentados anteriormente, porém representa um censo das pessoas que trabalham na organização. A base de dados encontra-se no arquivo **Satisfação.sav** e parte dela é apresentada na Figura 11.2. Para fins didáticos, serão apresentadas todas as diferenças entre os dois procedimentos disponíveis da análise discriminante e, para tanto, será utilizado o mesmo problema de pesquisa e as mesmas variáveis para os procedimentos simultâneo e *stepwise*.

Primeiramente, clique em **Analyze → Classify → Discriminant**, conforme mostra a Figura 11.3.

	nota	satisfa_emg	sal_atual	sal_inicial	tempo_ser	nivel_educ	idade
1	1	Pouco Satisfeito	29100	16500	72	14	24
2	1	Pouco Satisfeito	31350	13500	70	15	26
3	2	Pouco Satisfeito	19200	16800	62	10	24
4	2	Pouco Satisfeito	31350	12000	75	15	33
5	2	Pouco Satisfeito	36000	14250	65	16	25
6	4	Pouco Satisfeito	23550	13500	75	10	32
7	4	Pouco Satisfeito	35100	15000	73	15	29
8	5	Pouco Satisfeito	22350	11550	71	10	27
9	5	Pouco Satisfeito	23250	14250	64	10	24
10	5	Pouco Satisfeito	29250	27510	74	14	27
11	5	Pouco Satisfeito	30750	14250	72	14	28
12	7	Pouco Satisfeito	30000	15000	76	14	35
13	8	Satisfeito	33900	9000	82	15	37
14	8	Satisfeito	34800	9000	85	15	47
15	8	Satisfeito	35550	11550	74	15	31
16	9	Satisfeito	25050	14250	84	12	38
17	9	Satisfeito	27000	9500	86	12	48
18	10	Satisfeito	26400	10500	87	12	53
19	10	Satisfeito	26850	14250	78	12	34
20	10	Satisfeito	28050	15000	91	12	46
21	10	Satisfeito	33900	14250	75	15	34
22	12	Satisfeito	22500	9000	92	10	51
23	12	Satisfeito	30900	14100	90	14	47

Figura 11.2: Base de dados no SPSS.

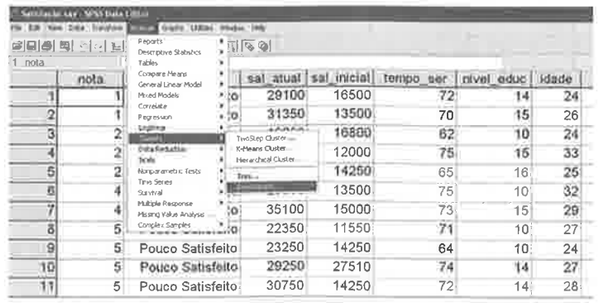


Figura 11.3: Procedimento para elaboração da AD no SPSS.

Uma janela como a apresentada na Figura 11.4 aparecerá. Insira a variável *satisfa_emg* na caixa **Grouping Variable** e defina a amplitude dos grupos, clicando em **Define Range**. Digite 1 e 3 em **Minimum** e **Maximum**, respectivamente, como demonstra a Figura 11.4, a fim de que todas as categorias da variável dependente sejam selecionadas para a formação dos grupos (no caso, portanto, 3 grupos). A escolha da variável dependente *satisfa_emg* propicia o início da modelagem AD, uma vez que, como ela é categórica, sua definição permite que o pesquisador estude o comportamento das variáveis explicativas para a formação dos grupos e a elaboração de previsões para a alocação de novos indivíduos em determinado grupo em função de suas características. Clique em **Continue**.

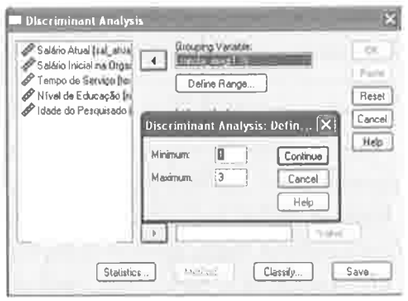


Figura 11.4: Seleção da variável dependente (*Grouping Variable*).

Na sequência, as variáveis discriminantes a serem inicialmente inseridas no modelo deverão ser escolhidas. Para tanto, selecione todas as variáveis (*sal_atual*, *sal_inicial*, *tempo_ser*, *nivel_educ* e *idade*) para inserção na caixa **Independents**, conforme mostrado por meio da Figura 11.5:

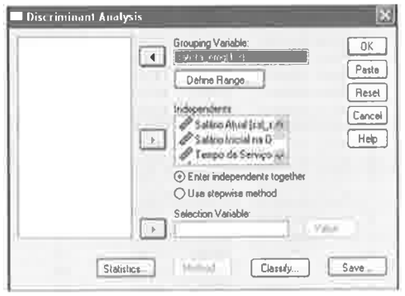


Figura 11.5: Escolha das variáveis explicativas.

Neste caso, como estamos aplicando o procedimento simultâneo, vamos deixar a primeira opção **Enter independents together** selecionada. A opção **Use stepwise method** será abordada na sequência.

A caixa **Selection Variable** pode ser utilizada quando o pesquisador deseja elaborar uma estratificação da amostra total em duas subamostras. Assim, uma parte seria destinada ao desenvolvimento da função discriminante e a outra para o teste da função elaborada. O tamanho da amostra pode eventualmente ser um obstáculo para a elaboração de estratificações. Ao final deste exemplo, apresentaremos um caso com a utilização de subamostras.

Clique em **Statistics** para selecionar as opções referentes às estatísticas que serão geradas para análise. Marque todas as opções, conforme apresentado na Figura 11.6. Estas estatísticas foram brevemente apresentadas na Seção 11.3 e a análise de seus *outputs* facilitará o entendimento da técnica. Clique em **Continue**.

Note que a opção **Method** não está disponível neste momento, uma vez que o procedimento de análise escolhido foi o simultâneo.

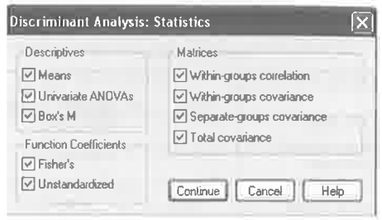


Figura 11.6: Seleção das estatísticas.

Em **Classify**, selecione a opção **Compute from group sizes** em **Prior Probabilities** e clique na opção **Summary table** em **Display**. Como neste exemplo utilizaremos para análise a matriz de covariância para todos os grupos, selecione a opção **Within-groups** em **Use Covariance Matrix**. Por fim, em **Plots**, selecione **Combined-groups** e **Territorial map**, de acordo com a Figura 11.7:

Clique em **Continue** e em **OK**.

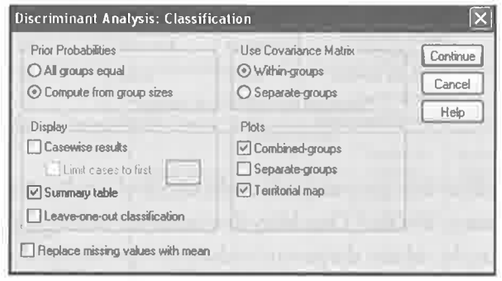


Figura 11.7: Opções para a análise da classificação.

11.5.2. Análise dos Resultados

A primeira tabela gerada nos *outputs* informa que, no nosso exemplo, todas as observações foram consideradas na análise.

Tabela 11.1: Observações Consideradas na AD

Analysis Case Processing Summary

Unweighted Cases		N	Percent
Valid		32	100,0
Excluded	Missing or out-of-range group codes	0	,0
	At least one missing discriminating variable	0	,0
	Both missing or out-of-range group codes and at least one missing discriminating variable	0	,0
	Total	0	,0
Total		32	100,0

A análise descritiva das variáveis é apresentada na Tabela 11.2, que mostra as médias, os desvios padrão e o número de observações em cada grupo, com total de 32 observações.

Tabela 11.2: Estatísticas Descritivas das Variáveis para cada Grupo

Group Statistics

Classificação de Satisfação		Mean	Std. Deviation	Valid N (listwise) Unweighted	Weighted
Pouco Satisfeito	Salário Atual	28437,50	5227,729	12	12,000
	Salário Inicial na Organização	15342,50	4130,512	12	12,000
	Tempo de Serviço	70,75	4,654	12	12,000
	Anos de Estudo	13,08	2,353	12	12,000
	Idade do Pesquisado	27,83	3,738	12	12,000
Satisfeito	Salário Atual	29536,36	4462,572	11	11,000
	Salário Inicial na Organização	11854,55	2531,151	11	11,000
	Tempo de Serviço	84,00	6,197	11	11,000
	Anos de Estudo	13,09	1,758	11	11,000
	Idade do Pesquisado	42,36	7,698	11	11,000
Muito Satisfeito	Salário Atual	49352,78	5422,097	9	9,000
	Salário Inicial na Organização	15047,78	4974,401	9	9,000
	Tempo de Serviço	88,22	7,710	9	9,000
	Anos de Estudo	18,78	1,481	9	9,000
	Idade do Pesquisado	46,67	7,450	9	9,000
Total	Salário Atual	34697,66	10520,664	32	32,000
	Salário Inicial na Organização	14060,63	4141,530	32	32,000
	Tempo de Serviço	80,22	9,684	32	32,000
	Anos de Estudo	14,69	3,207	32	32,000
	Idade do Pesquisado	38,13	10,342	32	32,000

O teste de igualdade de médias dos grupos para cada variável explicativa é apresentado a seguir, por meio da Tabela 11.3, que mostra a *ANOVA One Way* das variáveis referentes a salário atual, salário inicial na organização, nível de educação, idade do pesquisado e tempo de serviço. Esta tabela identifica também as variáveis que são as melhores discriminantes dos níveis de satisfação (muito satisfeito, satisfeito e pouco satisfeito).

Tabela 11.3: Teste de Igualdade de Médias dos Grupos

Tests of Equality of Group Means

	Wilks' Lambda	F	df1	df2	Sig.
Salário Atual	,214	53,194	2	29	,000
Salário Inicial na Organização	,846	2,645	2	29	,088
Tempo de Serviço	,378	23,902	2	29	,000
Nível de Educação	,343	27,771	2	29	,000
Idade do Pesquisado	,359	25,892	2	29	,000

Como já abordado, o lambda de Wilks, que varia de 0 a 1, testa a existência de diferenças de médias entre os grupos para cada variável. É importante lembrar que valores elevados desta estatística indicam ausência de diferenças entre os grupos.

Podemos perceber que a variável referente ao salário atual é a que mais discrimina os grupos, ou seja, seu poder de diferenciação dos grupos é superior, se comparado às outras variáveis. Por outro lado, a variável que se refere ao salário inicial na organização apresenta um valor mais elevado (0,846), demonstrando ser a pior em termos de discriminação dos grupos. O Sig. *F* expressa as diferenças entre as médias, sendo que os valores mais próximos de 0 indicam médias mais distintas. Seu cálculo é elaborado por meio da relação entre a soma dos quadrados dos erros totais dentro dos grupos e da soma dos quadrados dos erros totais.

Considerando-se uma probabilidade de erro tipo I de 5% (nível de significância $\alpha = 0,05$), podemos afirmar que somente a variável *sal_inicial* não se mostrou uma possível discriminante dos grupos. Já as variáveis relativas a tempo de serviço, nível de educação e idade do pesquisado mostraram-se possíveis discriminantes, assim como a variável *sal_atual*.

Podemos assumir, portanto, que, para a maioria das variáveis explicativas consideradas no exemplo, existe pelo menos um grupo em que as médias são diferentes.

Alguns pesquisadores não consideram o nível de significância para determinar a remoção da variável do teste. Mas caso a hipótese de igualdade de média seja verificada (Sig. $F > 0,05$) para determinada variável, alguma classificação poderá ser induzida ao erro. Não faremos a exclusão da variável *sal_inicial* agora, uma vez que o procedimento *stepwise*, a ser realizado adiante, se encarregará de excluí-la.

As próximas duas tabelas apresentam as matrizes de covariância e de correlação. Estas tabelas contribuem para a avaliação da relação entre as variáveis, e é a partir delas que podemos notar a presença de multicolinearidade entre os elementos. Caso ocorram correlações muito elevadas entre duas variáveis, recomenda-se a exclusão de uma delas ou a transformação de ambas para um fator que explique sua variância, conforme discutido no Capítulo 7 (Análise Fatorial).

Tabela 11.4: Matrizes de Covariância e de Correlação para todos os Grupos

Pooled Within-Groups Matrices[a]

		Salário Atual	Salário Inicial na Organização	Tempo de Serviço	Anos de Estudo	Idade do Pesquisado
Covariance	Salário Atual	25343418,5	4173615,944	-3554,933	9116,610	-4709,039
	Salário Inicial na Organização	4173615,944	15506794,941	-1904,416	1835,121	-7176,891
	Tempo de Serviço	-3554,933	-1904,416	37,855	-1,148	33,799
	Anos de Estudo	9116,610	1835,121	-1,148	3,772	-1,892
	Idade do Pesquisado	-4709,039	-7176,891	33,799	-1,892	41,042
Correlation	Salário Atual	1,000	,211	-,115	,932	-,146
	Salário Inicial na Organização	,211	1,000	-,079	,240	-,284
	Tempo de Serviço	-,115	-,079	1,000	-,096	,857
	Anos de Estudo	,932	,240	-,096	1,000	-,152
	Idade do Pesquisado	-,146	-,284	,857	-,152	1,000

a. The covariance matrix has 29 degrees of freedom.

Pelo nosso exemplo, a maior correlação positiva ocorreu entre o nível de educação e o salário atual (0,932), o que induz à conclusão de que maiores salários atuais estão relacionados a maiores níveis de educação. A variável referente ao tempo de serviço também apresentou uma alta correlação positiva com a variável *idade*. Como veremos por meio do procedimento *stepwise*, apenas as variáveis *sal_atual* e *idade* serão boas discriminantes para a composição dos grupos, em função desses problemas de multicolinearidade.

As matrizes de covariância para cada um dos grupos auxiliam o pesquisador quanto à percepção de homogeneidade de covariância. É importante lembrar que a presença de homogeneidade das matrizes de covariância é um dos pressupostos da AD. Entretanto, será por meio da estatística Box's M que teremos condições de verificar se as diferentes dispersões observadas são ou não estatisticamente significativas. Este teste tem como hipótese nula que não há diferenças significativas entre os grupos, ou seja, que há homogeneidade das matrizes de covariância para os grupos em análise. O resultado do teste é apresentado na Tabela 11.6.

Tabela 11.5: Matrizes de Covariância para cada um dos Grupos

Covariance Matrices[a]

Classificação de Satisfação		Salário Atual	Salário Inicial na Organização	Tempo de Serviço	Anos de Estudo	Idade do Pesquisado
Pouco Satisfeito	Salário Atual	27329147,7	567443,182	8164,773	11746,591	4038,636
	Salário Inicial na Organização	567443,182	17061129,545	1487,045	1215,227	-3120,455
	Tempo de Serviço	8164,773	1487,045	21,659	3,477	13,591
	Anos de Estudo	11746,591	1215,227	3,477	5,538	1,470
	Idade do Pesquisado	4038,636	-3120,455	13,591	1,470	13,970
Satisfeito	Salário Atual	19914545,5	-734931,818	-15975,000	7681,364	-16929,545
	Salário Inicial na Organização	-734931,818	6406727,273	-2615,000	-65,455	-7546,818
	Tempo de Serviço	-15975,000	-2615,000	38,400	-6,400	42,000
	Anos de Estudo	7681,364	-65,455	-6,400	3,091	-6,936
	Idade do Pesquisado	-16929,545	-7546,818	42,000	-6,936	59,255
Muito Satisfeito	Salário Atual	29399131,9	15267788,194	-4144,444	7294,444	-1461,458
	Salário Inicial na Organização	15267788,2	24744669,444	-5679,444	5063,194	-12292,083
	Tempo de Serviço	-4144,444	-5679,444	59,444	-,944	51,333
	Anos de Estudo	7294,444	5063,194	-,944	2,194	-,208
	Idade do Pesquisado	-1461,458	-12292,083	51,333	-,208	55,500
Total	Salário Atual	110684369	9038324,093	46747,303	32742,792	49114,012
	Salário Inicial na Organização	9038324,093	17152270,565	-7146,270	3342,782	-12690,726
	Tempo de Serviço	46747,303	-7146,270	93,789	12,167	94,875
	Anos de Estudo	32742,792	3342,782	12,167	10,286	12,363
	Idade do Pesquisado	49114,012	-12690,726	94,875	12,363	106,952

a. The total covariance matrix has 31 degrees of freedom.

Tabela 11.6: Resultado do Teste Box's M

Test Results

Box's M		45,361
F	Approx.	1,114
	df1	30
	df2	2331,929
	Sig.	,306

Tests null hypothesis of equal population covariance matrices.

Uma vez que este teste apresentou um Sig. *F* igual a 0,306, o que não permite a rejeição da hipótese nula a 5%, podemos concluir que não há significância das diferenças observadas, ou seja, que há igualdade das dispersões entre os grupos. Reforçamos que a estatística Box's M é muito influenciada pelo tamanho da amostra e pelas diferenças de tamanho que determinadas amostras podem apresentar. Ademais, é uma estatística que se mostra muito sensível à quebra do pressuposto de normalidade multivariada.

As próximas tabelas apresentam o sumário das funções discriminantes canônicas. A Tabela 11.7 mostra os *eigenvalues* para cada função discriminante:

Tabela 11.7: *Eigenvalues*

Eigenvalues

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	6,728[a]	90,4	90,4	,933
2	,716[a]	9,6	100,0	,646

a. First 2 canonical discriminant functions were used in the analysis.

Os *eigenvalues* representam o percentual de variância explicada em termos de diferenças entre os grupos e é uma medida relativa de quão diferentes os grupos são na função discriminante, ou seja, quanto mais afastados de 1 forem os *eigenvalues*, maiores serão as variações entre os grupos explicadas pela função discriminante (PESTANA e GAGEIRO, 2003).

Pelo *output* apresentado, a primeira função discriminante apresenta um percentual de 90,4% [6,728 / (6,728 + 0,716)], ou seja, esta função contribui mais para demonstrar as diferenças entre os grupos. Já a segunda função não demonstra um poder discriminante substancial, já que explica somente 9,6% [0,716 / (6,728 + 0,716)] da variância entre os grupos.

Neste exemplo, como há três grupos, duas funções discriminantes são definidas, sendo que a primeira discrimina os grupos de forma substancialmente melhor do que a segunda.

Ainda por meio da mesma tabela, podemos observar a coluna que apresenta as correlações canônicas, que correspondem à razão entre a variação entre os grupos e a variação total. Analogamente aos conceitos estudados no Capítulo 10 (Regressão Múltipla), quando há a presença de somente dois grupos, a correlação canônica passa a ser igual ao coeficiente de determinação R^2.

Neste caso, para cada *eigenvalue*, podemos calcular os respectivos valores de lambda de Wilks, por meio da Expressão (11.9). Assim, temos:

Função 1 para 2:

$$\Lambda_1 = \frac{1}{(1+6{,}728)} \cdot \frac{1}{(1+0{,}716)} = 0{,}075$$

Função 2:

$$\Lambda_2 = \frac{1}{(1+0{,}716)} = 0{,}583$$

Esses resultados estão apresentados na Tabela 11.8. Na primeira linha, são testadas as duas funções em conjunto, podendo-se concluir que pelo menos a primeira função discriminante é altamente significativa. A linha seguinte é referente à segunda função discriminante, sendo também possível rejeitar H_0 de que as médias dos grupos nesta função são iguais. Em outras palavras, podemos verificar, por meio da segunda linha, que o Sig. χ^2 (0,006) demonstra um decréscimo no poder discriminante por conta de um aumento no lambda de Wilks, embora ainda seja um valor significante em relação aos níveis habituais de significância.

Tabela 11.8: Lambda de Wilks e Qui-quadrado

Wilks' Lambda

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1 through 2	,075	69,790	10	,000
2	,583	14,580	4	,006

Voltando aos resultados da Tabela 11.7, podemos calcular as correlações canônicas por meio da Expressão (11.12). Assim, para cada lambda de Wilks, temos:

Função 1: 0,075 + $(CANCOR_1)^2$ = 1
$CANCOR_1$ = 0,933

Função 2: 0,583 + $(CANCOR_2)^2$ = 1
$CANCOR_2$ = 0,646

Da mesma forma, por meio da Expressão (11.10) podemos calcular os valores de Qui-quadrado apresentados na Tabela 11.8, que testam a significância das funções discriminantes, ou seja, o quão bem cada função separa as observações em grupos (testa se as médias são diferentes entre os grupos). Desta forma:

Função 1:

$$\chi_1^2 = -\left[32 - \frac{(5+3)}{2} - 1\right] \cdot \ln(0{,}075) = 69{,}790$$

Função 2:

$$\chi_1^2 = -\left[32 - \frac{(5+3)}{2} - 1\right] \cdot \ln(0{,}583) = 14{,}580$$

Portanto, é possível afirmar que as duas funções discriminantes são significantes para separar as observações em grupos. Essas estatísticas são importantes para verificarmos se as funções discriminantes refletem as diferenças entre os grupos.

A Tabela 11.9 apresenta os coeficientes não padronizados das funções discriminantes para cada uma das variáveis explicativas.

Tabela 11.9: Coeficientes das Funções Discriminantes

Canonical Discriminant Function Coefficients

	Function	
	1	2
Salário Atual	,00037803	-,000142
Salário Inicial na Organização	,00000340	-,000079
Tempo de Serviço	,04132696	,07140348
Anos de Estudo	-,61458368	,19101708
Idade do Pesquisado	,04947563	,05208704
(Constant)	-9,339123	-4,46796

Unstandardized coefficients

Por meio da Tabela 11.9, é possível escrevermos cada função da seguinte forma:

Z_1 = –9,339123 + 0,00037803.*sal_atual* + 0,00000340.*sal_inicial* + 0,04132696.*tempo_ser* – 0,61458368.*nivel_educ* + 0,04947563.*idade*

Z_2 = –4,46796 – 0,000142.*sal_atual* – 0,000079.*sal_inicial* + 0,07140348.*tempo_ser* + 0,19101708.*nivel_educ* + 0,05208704.*idade*

Os coeficientes padronizados das funções discriminantes são obtidos pela multiplicação dos coeficientes não padronizados pelas respectivas raízes das covariâncias para cada variável. Assim, a partir das Tabelas 11.4 e 11.9, temos:

Função 1:

$$W_1 = 0{,}00037803 \cdot \sqrt{25343418{,}5} = 1{,}903$$

$$W_2 = 0{,}00000340 \cdot \sqrt{15506794{,}941} = 0{,}013$$

$$W_3 = 0{,}04132696 \cdot \sqrt{37{,}855} = 0{,}254$$

$$W_4 = 0{,}61458368 \cdot \sqrt{3{,}772} = -1{,}194$$

$$W_5 = 0{,}04947563 \cdot \sqrt{41{,}042} = 0{,}317$$

Função 2:

$$W_1 = 0{,}000142 \cdot \sqrt{25343418{,}5} = 0{,}717$$

$$W_2 = 0{,}000079 \cdot \sqrt{15506794{,}941} = -0{,}310$$

$$W_3 = 0{,}07140348 \cdot \sqrt{37{,}855} = 0{,}439$$

$$W_4 = 0{,}19101708 \cdot \sqrt{3{,}772} = 0{,}371$$

$$W_5 = 0{,}05208704 \cdot \sqrt{41{,}042} = 0{,}334$$

Os valores dos coeficientes padronizados das funções discriminantes, que acabamos de calcular, são apresentados na Tabela 11.10 a seguir:

Tabela 11.10: Coeficientes Padronizados das Funções Discriminantes

Standardized Canonical Discriminant Function Coefficients

	Function	
	1	2
Salário Atual	1,903	-,717
Salário Inicial na Organização	,013	-,310
Tempo de Serviço	,254	,439
Nível de Educação	-1,194	,371
Idade do Pesquisado	,317	,334

Segundo Maroco (2007), esses coeficientes, que também são chamados de pesos discriminantes, podem ser utilizados para avaliar a importância relativa de cada variável explicativa para a função discriminante, sendo que a interpretação destas funções a partir deles deve ser feita com alguma precaução caso haja problemas de multicolinearidade.

Hair, Anderson, Tatham e Black (2005) definem peso discriminante como o parâmetro cuja magnitude é determinada pela estrutura de variância das variáveis originais nos grupos definidos pelas categorias da variável dependente. Assim, variáveis explicativas com grande poder discriminante geralmente apresentam grandes pesos, porém a presença de multicolinearidade pode gerar certa igualdade na magnitude dos pesos discriminantes.

Cargas discriminantes, por outro lado, referem-se à correlação linear simples entre as variáveis explicativas e os escores discriminantes para cada função discriminante. Também chamadas de correlações estruturais, as cargas discriminantes são calculadas para propiciar uma análise da hierarquia do poder discriminante das variáveis explicativas, de forma análoga à obtida pelos lambdas de Wilks apresentados na Tabela 11.3.

Quando a correlação não é muito forte, alguns autores, como Sharma (1996), sugerem que as correlações existentes entre as variáveis explicativas e as funções discriminantes sejam ponderadas pelos respectivos coeficientes padronizados (pesos), a fim de que sejam definidas as importâncias relativas das variáveis explicativas nas funções discriminantes. Assim, temos que:

$$\iota_{im} = \sum_{j=1}^{p} r_{ij} w_{jm}^{*} \qquad (11.23)$$

em que ι_{im} é a correlação estrutural da variável i com a função discriminante m, r_{ij} é a correlação conjunta entre grupos da variável i com a variável j ($j = 1, \ldots, p$) e w_{jm}^{*} representa o coeficiente padronizado da variável i com a função discriminante m.

Assim, a matriz de estrutura apresentada a seguir, na Tabela 11.11, auxilia na interpretação da contribuição que cada variável forneceu para cada função discriminante, uma vez que apresenta as correlações entre as variáveis explicativas e as funções discriminantes canônicas padronizadas.

Tabela 11.11: Matriz de Estrutura

Structure Matrix

	Function	
	1	2
Salário Atual	,717*	-,536
Nível de Educação	,512*	-,465
Idade do Pesquisado	,435	,847*
Tempo de Serviço	,421	,797*
Salário Inicial na Organização	,017	-,502*

Pooled within-groups correlations between discriminating variables and standardized canonical discriminant functions
Variables ordered by absolute size of correlation within function.

*. Largest absolute correlation between each variable and any discriminant function

As variáveis cujos valores apresentam-se com asterisco são as mais relevantes para a determinação de cada função discriminante, uma vez que oferecem maiores correlações com essas funções. Assim, enquanto a primeira função discriminante tem maior correlação com as variáveis *sal_atual* e *nivel_educ*, a segunda função apresenta maior correlação com as variáveis *idade*, *tempo_ser* e *sal_inicial*. Quando da elaboração do método *stepwise* mais adiante, será possível verificar que apenas as variáveis com maior correlação com cada função canônica serão incluídas no modelo final, ou seja, *sal_atual* na primeira função e *idade* na segunda função discriminante.

A partir dos coeficientes não padronizados das funções discriminantes, é possível definir a posição de cada um dos centróides dos grupos em um mapa territorial. Estas coordenadas encontram-se a seguir, na Tabela 11.12:

Tabela 11.12: Centróides dos Grupos

Functions at Group Centroids

Classificação de Satisfação	Function	
	1	2
Pouco Satisfeito	-2,277	-,728
Satisfeito	-,611	1,095
Muito Satisfeito	3,783	-,368

Unstandardized canonical discriminant functions evaluated at group means

As Tabelas 11.13 e 11.14 apresentam, respectivamente, um sumário das observações utilizadas na análise e as probabilidades calculadas *a priori* a partir da amostra para a obtenção do ponto de corte crítico (MAROCO, 2007). Essas probabilidades foram calculadas porque selecionamos a opção **Compute from group sizes** em **Prior Probabilities** no menu **Classify**.

Tabela 11.13: Sumário das Observações Utilizadas na Análise

Classification Processing Summary

Processed		32
Excluded	Missing or out-of-range group codes	0
	At least one missing discriminating variable	0
Used in Output		32

Tabela 11.14: Probabilidades Calculadas *a Priori*

Prior Probabilities for Groups

Classificação de Satisfação	Prior	Cases Used in Analysis	
		Unweighted	Weighted
Pouco Satisfeito	,375	12	12,000
Satisfeito	,344	11	11,000
Muito Satisfeito	,281	9	9,000
Total	1,000	32	32,000

A definição do ponto de corte auxilia na classificação de novos elementos. A Tabela 11.15 apresenta os coeficientes das funções de classificação, que servem apenas para classificar observações e não têm qualquer interpretação discriminante (MAROCO, 2007).

Tabela 11.15: Coeficientes de Classificação das Funções Discriminantes

Classification Function Coefficients

	Classificação de Satisfação		
	Pouco Satisfeito	Satisfeito	Muito Satisfeito
Salário Atual	3,17E-006	,00037305	,00224258
Salário Inicial na Organização	-,0002834	-,000421	-,0002912
Tempo de Serviço	4,7436930	4,942674	5,0198301
Anos de Estudo	3,4798717	2,804554	-,1755152
Idade do Pesquisado	-3,117150	-2,93981	-2,798590
(Constant)	-146,0440	-167,758	-208,8985

Fisher's linear discriminant functions

Assim, cada observação será classificada no grupo em que o escore discriminante for maior. Por exemplo, imagine um novo funcionário da empresa que tem as seguintes características:

- Salário atual: $ 29.000,00
- Salário inicial: $ 29.000,00
- Tempo de serviço: 82 meses
- Anos de estudo: 25 anos
- Idade: 31 anos

Qual o possível grupo a que pertenceria esse profissional?
Para tanto, devemos calcular os escores dos três grupos, da seguinte maneira:

Pouco Satisfeito:

$$0{,}00000.(29000) - 0{,}00028.(29000) + 4{,}74369.(82) + 3{,}47987.(25) - 3{,}11715.(31) - 146{,}0440 = \mathbf{225{,}18368}$$

Satisfeito:

$$0{,}00037.(29000) - 0{,}00042.(29000) + 4{,}94267.(82) + 2{,}80455.(25) - 2{,}93981.(31) - 167{,}758 = \mathbf{215{,}07058}$$

Muito Satisfeito:

$$0{,}00224.(29000) - 0{,}00029.(29000) + 5{,}01983.(82) - 0{,}17551.(25) - 2{,}79859.(31) - 208{,}8985 = \mathbf{168{,}13352}$$

Dessa forma, este novo funcionário, com as características mencionadas, pertenceria ao grupo dos pouco satisfeitos, uma vez que é neste grupo que se observa o maior valor das funções de classificação. Assim sendo, a empresa em questão poderia facilmente definir um eventual incremento salarial para, por exemplo, tornar este funcionário satisfeito ou até muito satisfeito, mantidas as demais variáveis constantes.

No procedimento simultâneo, os escores discriminantes dos grupos são calculados levando-se em consideração a inclusão de todas as variáveis explicativas. No procedimento *stepwise*, a ser realizado adiante, possivelmente uma ou mais variáveis serão excluídas pela existência de multicolinearidade e, portanto, os escores discriminantes serão calculados somente com a consideração das variáveis incluídas no modelo final.

O mapa territorial do nosso exemplo é apresentado na Figura 11.8 a seguir, com destaque para os centróides de cada grupo. As zonas de fronteira de cada par de grupos são definidas pela Expressão (11.17).

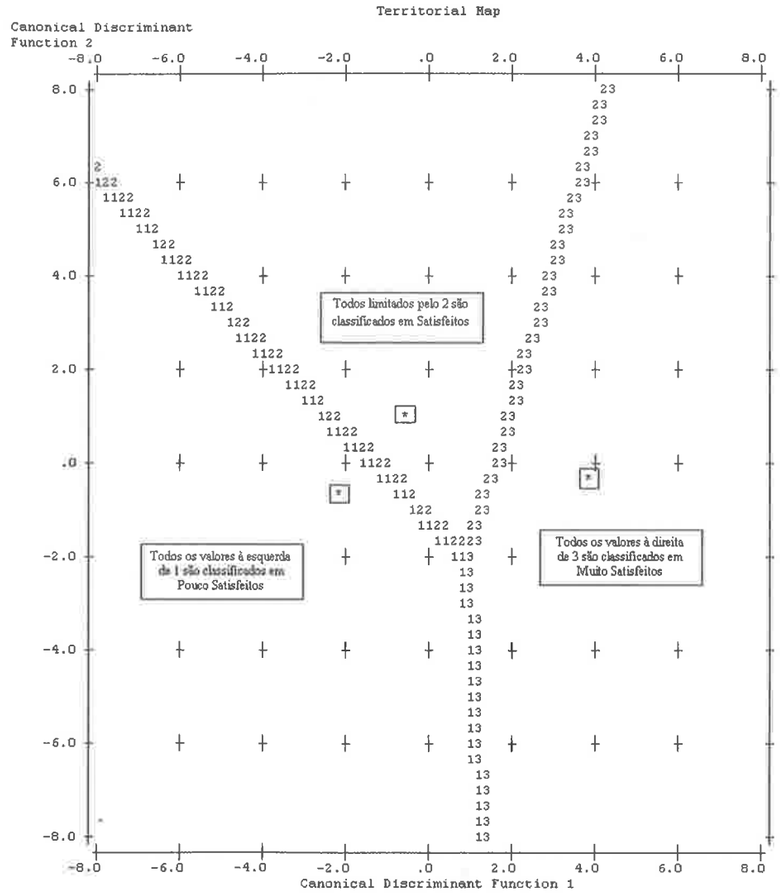


Figura 11.8: Mapa territorial da análise discriminante.

Em termos gerais, a classificação ocorre de modo que o caso seja classificado no grupo cujo centróide se encontra mais próximo.

A Figura 11.9, na sequência, apresenta a representação gráfica dos centróides de cada grupo nas duas funções discriminantes.

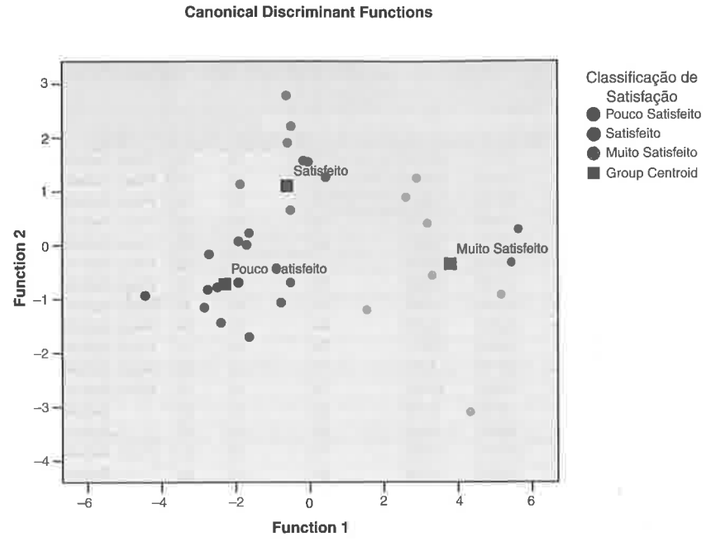


Figura 11.9: Representação gráfica dos centróides nas funções discriminantes.

Por fim, a Tabela 11.16 apresenta os resultados da classificação.

Tabela 11.16: Resultados da Classificação

Classification Results[a]

			Predicted Group Membership			
		Classificação de Satisfação	Pouco Satisfeito	Satisfeito	Muito Satisfeito	Total
Original	Count	Pouco Satisfeito	12	0	0	12
		Satisfeito	3	8	0	11
		Muito Satisfeito	0	0	9	9
	%	Pouco Satisfeito	100,0	,0	,0	100,0
		Satisfeito	27,3	72,7	,0	100,0
		Muito Satisfeito	,0	,0	100,0	100,0

a. 90,6% of original grouped cases correctly classified.

Podemos perceber que 90,6% das observações foram classificadas corretamente e que apenas três funcionários satisfeitos foram classificados de forma errada no grupo de pouco satisfeito.

No menu **Analyze → Classify → Discriminant**, clique em **Save** e selecione a opção **Predicted group membership**, de acordo com a Figura 11.10. Clique em **Continue** e em **OK**.

Figura 11.10: *Predicted Group Membership*.

Esta opção faz com que uma nova variável (*Dis_1*) seja incluída no banco de dados, com os resultados dos grupos preditos. Este procedimento, realizado para cada observação, foi elaborado da mesma forma que o desenvolvido anteriormente quando do cálculo dos escores das funções de classificação para a definição do grupo do novo funcionário. Note que os próprios escores discriminantes e as probabilidades de pertencer a determinado grupo podem ser solicitados. A Figura 11.11 apresenta este novo banco de dados, por meio do qual é possível verificar quais foram as três observações classificadas de modo errado.

	nota	satisfa_emg	sal_atual	sal_inicial	tempo_ser	nivel_educ	idade	Dis_1	var	var
1	1	Pouco Satisfeito	29100	16500	72	14	24	Pouco Satisfeito		
2	1	Pouco Satisfeito	31350	13500	70	15	26	Pouco Satisfeito		
3	2	Pouco Satisfeito	19200	16800	62	10	24	Pouco Satisfeito		
4	2	Pouco Satisfeito	31350	12000	75	15	33	Pouco Satisfeito		
5	2	Pouco Satisfeito	36000	14250	65	16	25	Pouco Satisfeito		
6	4	Pouco Satisfeito	23550	13500	75	10	32	Pouco Satisfeito		
7	4	Pouco Satisfeito	35100	15000	73	15	29	Pouco Satisfeito		
8	5	Pouco Satisfeito	22350	11550	71	10	27	Pouco Satisfeito		
9	5	Pouco Satisfeito	23250	14250	64	10	24	Pouco Satisfeito		
10	5	Pouco Satisfeito	29250	27510	74	14	27	Pouco Satisfeito		
11	5	Pouco Satisfeito	30750	14250	72	14	28	Pouco Satisfeito		
12	7	Pouco Satisfeito	30000	15000	76	14	35	Pouco Satisfeito		
13	8	Satisfeito	33900	9000	82	15	37	Satisfeito		
14	8	Satisfeito	34800	9000	85	15	47	Satisfeito		
15	8	Satisfeito	35550	11550	74	15	31	Pouco Satisfeito		
16	9	Satisfeito	25050	14250	84	12	38	Satisfeito		
17	9	Satisfeito	27000	9500	86	12	48	Satisfeito		
18	10	Satisfeito	26400	10500	87	12	53	Satisfeito		
19	10	Satisfeito	26850	14250	78	12	34	Pouco Satisfeito		
20	10	Satisfeito	28050	15000	91	12	46	Satisfeito		
21	10	Satisfeito	33900	14250	75	15	34	Pouco Satisfeito		
22	12	Satisfeito	22500	9000	92	10	51	Satisfeito		
23	12	Satisfeito	30900	14100	90	14	47	Satisfeito		
24	15	Muito Satisfeito	45150	12600	76	18	35	Muito Satisfeito		
25	15	Muito Satisfeito	55000	13500	85	20	49	Muito Satisfeito		
26	16	Muito Satisfeito	53125	15000	94	19	49	Muito Satisfeito		
27	17	Muito Satisfeito	48000	11250	84	18	42	Muito Satisfeito		
28	17	Muito Satisfeito	54000	15000	97	20	56	Muito Satisfeito		

Figura 11.11: Banco de dados com a variável *Dis_1* (*Predicted Group Membership*).

11.5.3. Exemplo com Subamostras

Para a elaboração de subamostras com disposições aleatórias, clique em **Transform, Compute Variable** e crie uma nova variável com o nome *disp_aleat*. Insira a função **UNIFORM** em **Numeric Expression**, de modo a fazer com que 70% das observações recebam uma distribuição uniforme de valor 0 e 30% o valor 1, como apresentado por meio da Figura 11.12.

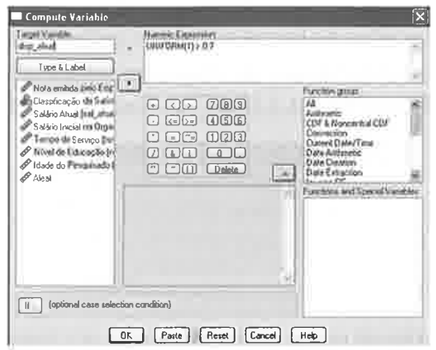


Figura 11.12: Criação de uma subamostra.

Por meio dos mesmos procedimentos já realizados para a elaboração da AD, insira, em **Selection Variable**, a nova variável *disp_aleat* e clique em **Value** para escolher o valor 0. Na sequência, clique em **Continue** para gerar a Figura 11.13.

Por fim, clique em **OK**.

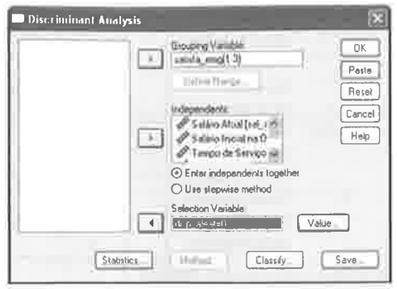


Figura 11.13: Procedimento AD com subamostra.

Por meio da Tabela 11.17, é possível verificarmos que 81,25% (26 casos) da amostra inicial formam a subamostra aleatória e 18,75% (6 casos) formam a amostra de teste, que, como já mencionado, terá como responsabilidade testar a validade das funções discriminantes elaboradas pela outra subamostra.

Tabela 11.17: Dimensões das Subamostras

Analysis Case Processing Summary

Unweighted Cases		N	Percent
Valid		26	81,3
Excluded	Missing or out-of-range group codes	0	,0
	At least one missing discriminating variable	0	,0
	Both missing or out-of-range group codes and at least one missing discriminating variable	0	,0
	Unselected	6	18,8
	Total	6	18,8
Total		32	100,0

Não iremos aqui analisar novamente todos os *outputs* gerados. Porém, a tabela de resultados da classificação (Tabela 11.18) apresenta algumas informações importantes.

Enquanto 96,2% das observações da amostra selecionada foram corretamente classificadas, com apenas um funcionário satisfeito sendo classificado erroneamente como pouco satisfeito, na amostra de teste apenas 50% foi classificado corretamente, com três funcionários satisfeitos sendo classificados como pouco satisfeitos. Na verdade, este procedimento precisa ser realizado diversas vezes, a fim de que possamos de fato concluir se a amostra de teste é relevante ou não para testar a validade das funções discriminantes elaboradas pela outra subamostra, uma vez que diversas aplicações deste procedimento oferecerão resultados de adequação diferentes, já que a variável *disp_aleat* foi gerada de maneira aleatória. Ademais, como a amostra inicial não apresenta um grande tamanho, qualquer subamostra resultante terá pequenas dimensões.

Tabela 11.18: Resultados da Classificação para cada Subamostra

Classification Results[a,b]

			Classificação de Satisfação	Predicted Group Membership: Pouco Satisfeito	Satisfeito	Muito Satisfeito	Total
Cases Selected	Original	Count	Pouco Satisfeito	12	0	0	12
			Satisfeito	1	6	0	7
			Muito Satisfeito	0	0	7	7
		%	Pouco Satisfeito	100,0	,0	,0	100,0
			Satisfeito	14,3	85,7	,0	100,0
			Muito Satisfeito	,0	,0	100,0	100,0
Cases Not Selected	Original	Count	Pouco Satisfeito	0	0	0	0
			Satisfeito	3	1	0	4
			Muito Satisfeito	0	0	2	2
		%	Pouco Satisfeito	,0	,0	,0	100,0
			Satisfeito	75,0	25,0	,0	100,0
			Muito Satisfeito	,0	,0	100,0	100,0

a. 96,2% of selected original grouped cases correctly classified.

b. 50,0% of unselected original grouped cases correctly classified.

11.6. ANÁLISE DISCRIMINANTE E PROCEDIMENTO *STEPWISE*: UM EXEMPLO PRÁTICO

Para o procedimento *stepwise*, consideramos o mesmo exemplo utilizado até agora.

Após definirmos as variáveis explicativas e a variável dependente, devemos selecionar a opção **Use stepwise method**, conforme apresentado na Figura 11.14 a seguir:

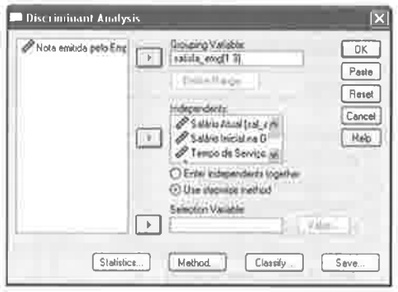


Figura 11.14: Seleção do procedimento *Stepwise* na AD.

Em **Method**, selecione o tipo de teste a ser aplicado quando da elaboração do procedimento *stepwise* na AD. Os possíveis testes já foram brevemente apresentados na Seção 11.5 e, no nosso exemplo, utilizaremos o método lambda de Wilks. Em **Display**, marque a opção **Summary of steps**, que propicia a apresentação das tabelas com as variáveis inseridas e removidas do modelo e os respectivos valores de lambda de Wilks. O critério de inclusão ou exclusão de variáveis será mantido de acordo com o padrão do *software*, porém é possível tornar tal critério mais ou menos flexível, em função das necessidades do pesquisador. Estas opções são apresentadas na Figura 11.15 a seguir:

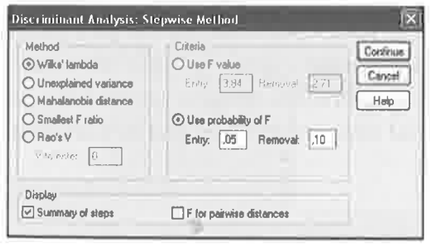


Figura 11.15: Opções do procedimento *Stepwise* na AD.

Clique em **Continue** e em **OK**, mantendo todas as demais opções conforme o já realizado quando da elaboração do procedimento simultâneo. Não discutiremos, entretanto, os *outputs* que forem iguais aos já fornecidos pelo procedimento simultâneo.

Assim, o resultado do teste Box's M é apresentado na Tabela 11.19 e, por meio desta, podemos concluir que não há significância das diferenças observadas, ou seja, que há igualdade das dispersões entre os grupos, uma vez que este teste apresentou um Sig. *F* igual a 0,245, o que não permite a rejeição da hipótese nula a 5%.

Tabela 11.19: Resultado do Teste Box's M para o Procedimento *Stepwise*

Test Results

Box's M		8,809
F	Approx.	1,318
	df1	6
	df2	13979,794
	Sig.	,245

Tests null hypothesis of equal population covariance matrices.

Analisaremos agora os resultados do procedimento *stepwise*, por meio das Tabelas 11.20, 11.21 e 11.22.

Tabela 11.20: Variáveis Incluídas/Excluídas no Modelo – Procedimento *Stepwise*

Variables Entered/Removed[a,b,c,d]

		Wilks' Lambda: Statistic	df1	df2	df3	Exact F: Statistic	df1	df2	Sig.
Step	Entered								
1	Salário Atual	,214	1	2	29,000	53,194	2	29,000	,000
2	Idade do Pesquisado	,095	2	2	29,000	31,539	4	56,000	,000

At each step, the variable that minimizes the overall Wilks' Lambda is entered.

a. Maximum number of steps is 10.

b. Maximum significance of F to enter is ,05.

c. Minimum significance of F to remove is ,10.

d. F level, tolerance, or VIN insufficient for further computation.

Tabela 11.21: Variáveis Discriminantes em cada Passo (*Stepwise*) da Análise

Variables in the Analysis

Step		Tolerance	Sig. of F to Remove	Wilks' Lambda
1	Salário Atual	1,000	,000	
2	Salário Atual	,979	,000	,359
	Idade do Pesquisado	,979	,000	,214

A Tabela 11.20 apresenta as variáveis incluídas / excluídas da análise por meio do procedimento *stepwise*, os respectivos valores de lambda de Wilks e a transformação para a estatística *F*. Segundo Maroco (2007), para cada passo do algoritmo escolhido (neste caso, o lambda de Wilks), a variável adicionada, dentre as possíveis variáveis explicativas, é aquela que minimiza o valor de lambda, ou seja, aquela para a qual ocorrem as maiores diferenças entre as médias dos grupos, até que não mais ocorram variações significativas de lambda.

Tabela 11.22: Variáveis Excluídas da Análise em cada Passo (*Stepwise*)

Variables Not in the Analysis

Step		Tolerance	Min. Tolerance	Sig. of F to Enter	Wilks' Lambda
0	Salário Atual	1,000	1,000	,000	,214
	Salário Inicial na Organização	1,000	1,000	,088	,846
	Tempo de Serviço	1,000	1,000	,000	,378
	Nível de Educação	1,000	1,000	,000	,343
	Idade do Pesquisado	1,000	1,000	,000	,359
1	Salário Inicial na Organização	,956	,956	,094	,181
	Tempo de Serviço	,987	,987	,000	,101
	Nível de Educação	,131	,131	,024	,164
	Idade do Pesquisado	,979	,979	,000	,095
2	Salário Inicial na Organização	,890	,890	,788	,093
	Tempo de Serviço	,265	,262	,816	,093
	Nível de Educação	,130	,130	,100	,080

A Tabela 11.21 fornece as variáveis discriminantes em cada passo da análise. Podemos perceber que apenas a variável *sal_atual* é incluída no modelo no passo 1. Já no passo 2, são incluídas as variáveis *sal_atual* e *idade*. Por mais que as variáveis *tempo_ser* e *nivel_educ* tenham se mostrado significantes na tabela de *ANOVA One Way* (Tabela 11.3), ou seja, apresentaram diferenças significativas nos três grupos de funcionários, elas não foram incluídas na análise. Como pode ser observado por meio da Tabela 11.23 a seguir, obtida por meio do procedimento **Analyze → Correlate → Bivariate**, que apresenta a matriz de correlações totais (não devemos confundir esta com a matriz de correlações entre os grupos, dada pela Tabela 11.4), as variáveis *sal_atual* e *nivel_educ* apresentam altas correlações. O mesmo pode ser dito em relação às variáveis *idade* e *tempo_ser*. Portanto, para cada um desses pares de variáveis, ficaram no modelo apenas as com menores lambdas de Wilks (maiores estatísticas *F*), conforme apresentado na Tabela 11.3. Ou seja, entre *sal_atual* e *nivel_educ*, escolheu-se a variável *sal_atual* e entre *idade* e *tempo_ser*, escolheu-se a variável *idade*.

Esses problemas de multicolinearidade são identificados pelo procedimento *stepwise* por meio do cálculo da Tolerância (*Tolerance*) de cada variável, conforme já apresentado no Capítulo 10 (Regressão Múltipla). Assim, segundo Maroco (2007), variáveis com maiores Tolerâncias devem ser incluídas no modelo, uma vez que uma grande proporção da variância de cada uma delas não é explicada pelas demais variáveis. A Tabela 11.22 apresenta os resultados da Tolerância para cada variável em cada passo e, por meio dessa tabela, podemos perceber que, em cada passo, apenas uma variável saiu da lista de excluídas (aquela com menor Sig. *F*, ou menor lambda de Wilks), até que mais nenhuma apresentasse um Sig. *F* menor do que 0,05 no passo 2.

Como era de se esperar, a variável *sal_inicial* também não foi incluída por não apresentar diferenças significativas nos três grupos (Sig. $F = 0{,}088 > 0{,}05$).

Seguindo a lógica proposta por Maroco (2007), as Tabelas 11.24 a 11.29 resumem a análise discriminante elaborada apenas com as variáveis selecionadas por meio do procedimento *stepwise*. A interpretação é exatamente a mesma daquela já realizada quando da elaboração deste exemplo por meio do procedimento simultâneo.

Tabela 11.23: Matriz de Correlações Totais das Variáveis Explicativas

Correlations

		Salário Atual	Salário Inicial na Organização	Tempo de Serviço	Anos de Estudo	Idade do Pesquisado
Salário Atual	Pearson Correlation	1	,207	,459**	,970**	,451**
	Sig. (2-tailed)		,255	,008	,000	,010
	N	32	32	32	32	32
Salário Inicial na Organização	Pearson Correlation	,207	1	-,178	,252	-,296
	Sig. (2-tailed)	,255		,329	,165	,100
	N	32	32	32	32	32
Tempo de Serviço	Pearson Correlation	,459**	-,178	1	,392*	,947**
	Sig. (2-tailed)	,008	,329		,027	,000
	N	32	32	32	32	32
Anos de Estudo	Pearson Correlation	,970**	,252	,392*	1	,373*
	Sig. (2-tailed)	,000	,165	,027		,036
	N	32	32	32	32	32
Idade do Pesquisado	Pearson Correlation	,451**	-,296	,947**	,373*	1
	Sig. (2-tailed)	,010	,100	,000	,036	
	N	32	32	32	32	32

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

Tabela 11.24: *Eigenvalues* – Procedimento *Stepwise*

Eigenvalues

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	5,466[a]	89,6	89,6	,919
2	,636[a]	10,4	100,0	,624

a. First 2 canonical discriminant functions were used in the analysis.

Tabela 11.25: Lambda de Wilks e Qui-quadrado – Procedimento *Stepwise*

Wilks' Lambda

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1 through 2	,095	67,232	4	,000
2	,611	14,035	1	,000

Tabela 11.26: Coeficientes das Funções Discriminantes – Procedimento *Stepwise*

Canonical Discriminant Function Coefficients

	Function	
	1	2
Salário Atual	,00018	-,00010
Idade do Pesquisado	,09626	,12502
(Constant)	-9,75162	-1,36786

Unstandardized coefficients

Tabela 11.27: Coeficientes Padronizados das Funções Discriminantes – Procedimento *Stepwise*

Standardized Canonical Discriminant Function Coefficients

	Function	
	1	2
Salário Atual	,882	-,493
Idade do Pesquisado	,617	,801

Tabela 11.28: Matriz de Estrutura – Procedimento *Stepwise*

Structure Matrix

	Function	
	1	2
Salário Atual	,792*	-,610
Nível de Educação[a]	,729*	-,582
Idade do Pesquisado	,488	,873*
Tempo de Serviço[a]	,427	,743*
Salário Inicial na Organização[a]	,010	-,332*

Pooled within-groups correlations between discriminating variables and standardized canonical discriminant functions
Variables ordered by absolute size of correlation within function.

*. Largest absolute correlation between each variable and any discriminant function

a. This variable not used in the analysis.

Tabela 11.29: Centróides dos Grupos – Procedimento *Stepwise*

Functions at Group Centroids

Classificação de Satisfação	Function	
	1	2
Pouco Satisfeito	-2,088	-,674
Satisfeito	-,497	1,035
Muito Satisfeito	3,391	-,368

Unstandardized canonical discriminant functions evaluated at group means

Por meio da análise dessas tabelas, é possível afirmar que a maior proporção de variância, em termos de diferenças entre os grupos, é explicada pela primeira função discriminante, porém ambas as funções são estatisticamente significantes.

A análise dos coeficientes padronizados permite dizer que, enquanto a variável *sal_atual* está alocada prioritariamente na primeira função discriminante, a variável *idade* define mais fortemente a segunda função discriminante. A mesma conclusão pode ser obtida por meio da análise da matriz de estrutura, já que as variáveis assinaladas com a letra *a* são aquelas que não entraram nas funções discriminantes.

A Tabela 11.30 apresenta os coeficientes de classificação das funções discriminantes.

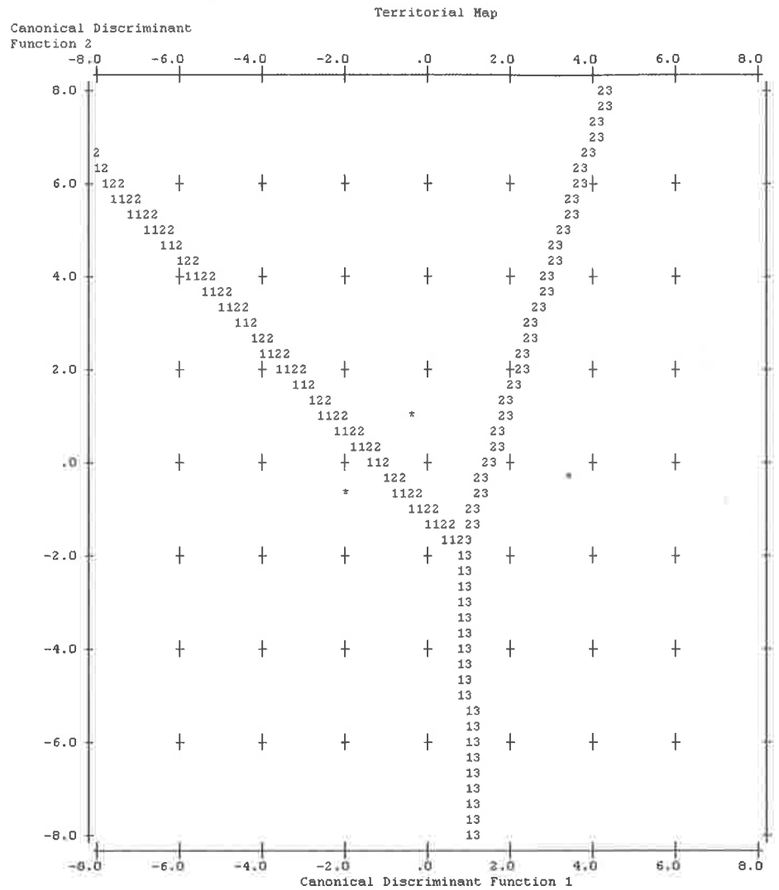
Tabela 11.30: Coeficientes de Classificação das Funções Discriminantes – Procedimento *Stepwise*

Classification Function Coefficients

	Classificação de Satisfação		
	Pouco Satisfeito	Satisfeito	Muito Satisfeito
Salário Atual	,001	,001	,002
Idade do Pesquisado	,824	1,191	1,390
(Constant)	-30,588	-46,783	-88,132

Fisher's linear discriminant functions

O mapa territorial e a representação gráfica dos centróides nas funções discriminantes são apresentados nas Figuras 11.16 e 11.17.



Symbols used in territorial map

Symbol	Group	Label
1	1	Pouco Satisfeito
2	2	Satisfeito
3	3	Muito Satisfeito
*		Indicates a group centroid

Figura 11.16: Mapa territorial da análise discriminante – procedimento *Stepwise*.

O principal objetivo da AD fica bastante explicitado com a análise das Figuras 11.18 e 11.19. Como apenas duas variáveis explicativas foram incluídas no modelo de AD após o procedimento *stepwise*, é possível elaborarmos um gráfico com a representação das observações para cada valor das variáveis *sal_atual* e *idade*, conforme a Figura 11.18. Por meio da análise dessa figura, é possível verificarmos que as duas variáveis apresentam diferenças de médias entre os grupos e homogeneidade de variância, porém sem que haja uma separação absoluta das distribuições.

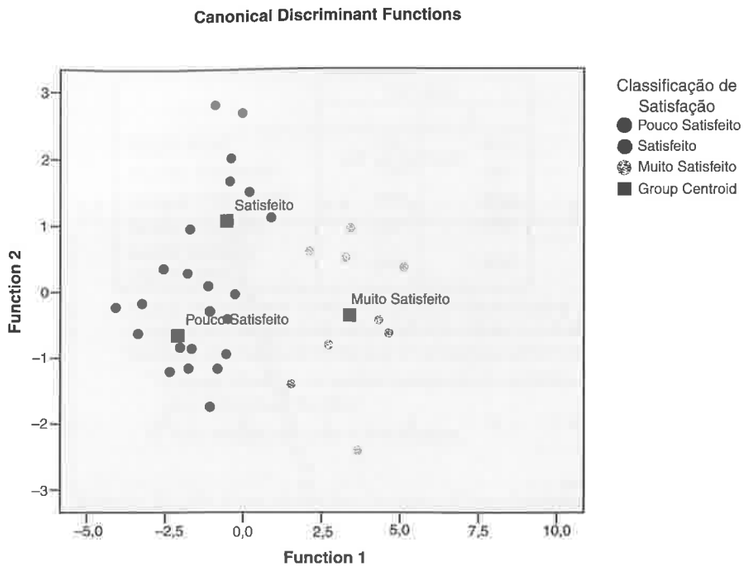


Figura 11.17: Representação gráfica dos centróides nas funções discriminantes – procedimento *Stepwise*.

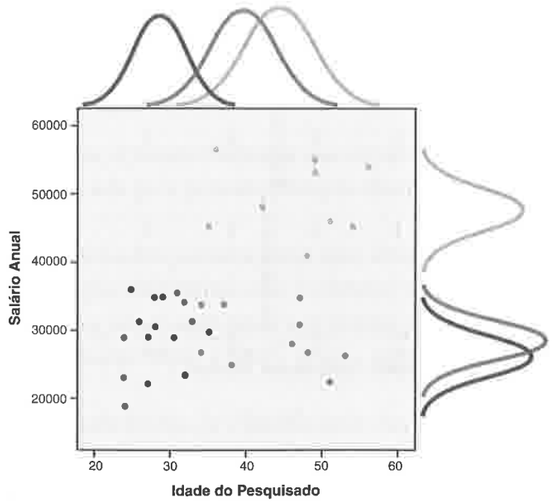


Figura 11.18: Representação gráfica das observações por grupo.

Neste sentido, são criadas duas funções discriminantes, Z_1 e Z_2, a fim de que sejam maximizadas as diferenças entre os grupos e minimizada a heterogeneidade dentro dos grupos. A Figura 11.19 mostra, de forma apenas ilustrativa, as funções discriminantes, cujos coeficientes são provenientes da Tabela 11.26.

Por fim, são apresentados na Tabela 11.31 os resultados da classificação após o procedimento *stepwise*. Podemos perceber que houve uma melhora do percentual de observações que foram classificadas corretamente (93,8%), já que, para este caso, apenas dois funcionários satisfeitos foram classificados de forma errada no grupo de pouco satisfeito.

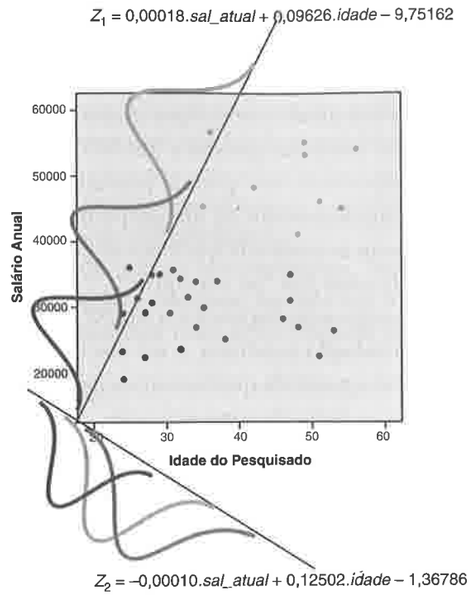


Figura 11.19: Criação das funções discriminantes.

Tabela 11.31: Resultados da Classificação – Procedimento *Stepwise*

Classification Results[a]

		Classificação de Satisfação	Predicted Group Membership: Pouco Satisfeito	Satisfeito	Muito Satisfeito	Total
Original	Count	Pouco Satisfeito	12	0	0	12
		Satisfeito	2	9	0	11
		Muito Satisfeito	0	0	9	9
	%	Pouco Satisfeito	100,0	,0	,0	100,0
		Satisfeito	18,2	81,8	,0	100,0
		Muito Satisfeito	,0	,0	100,0	100,0

a. 93,8% of original grouped cases correctly classified.

11.7. RELAÇÃO COM OUTRAS TÉCNICAS

A técnica de regressão múltipla é, sem dúvida, a mais utilizada quando o pesquisador está interessado nas relações de dependência de uma variável em relação ao comportamento de um conjunto de variáveis explicativas. Contudo, a usabilidade desta técnica restringe-se à utilização de variáveis dependentes métricas.

Embora a interpretação dos coeficientes gerados por meio da análise discriminante seja similar à da análise de regressão, os parâmetros de estimação utilizados pela AD, como a maximização da razão entre a variância entre os grupos e a variância dentro dos grupos são sensivelmente diferentes, uma vez que, na análise de regressão, há a preocupação com a minimização da soma dos quadrado dos resíduos (GOUVÊA, FARINA e VARELA, 2007).

A análise discriminante possui uma forte semelhança com a técnica de regressão logística, a ser estudada no Capítulo 12. Como ambas possuem caráter preditivo e classificatório, além de problemas de pesquisa por vezes similares, é frequente o pesquisador ficar com certa dúvida em relação à utilização de uma ou de outra. Porém, quando algumas das variáveis explicativas não forem métricas, pode ocorrer uma

"quebra" do pressuposto de normalidade multivariada na AD e, neste caso, a regressão logística oferece uma vantagem, por ser uma técnica com pressupostos bem mais flexíveis. Por outro lado, na regressão logística, a variável dependente é uma *dummy* e, portanto, o pesquisador pode estar interessado no cálculo das probabilidades de ocorrência de determinado evento que apresenta apenas duas possíveis categorias.

A análise discriminante também tem semelhança com a MANOVA. Como veremos no Capítulo 13, a MANOVA torna-se similar à análise discriminante quando possuir apenas uma variável explicativa não métrica (categórica) para avaliar o comportamento de um grupo de variáveis dependentes métricas, ou seja, na MANOVA, os grupos constituem as variáveis explicativas (MAROCO, 2007). Neste caso, dizemos que a MANOVA pode ser considerada uma AD "ao contrário".

Assim como a regressão múltipla, a regressão logística e a MANOVA, a análise discriminante também é um caso particular da técnica de correlação canônica, a ser discutida no Capítulo 14, que é a técnica multivariada de dependência mais geral a partir da qual as outras derivam, já que admite diversas variáveis métricas ou não métricas de ambos os lados de uma equação.

A análise discriminante possui também objetivos similares aos da análise fatorial, uma vez que ambas utilizam os cálculos dos *eigenvalues*. Porém, enquanto a análise fatorial estipula que a variância explicada por um novo fator criado seja máxima, a análise discriminante substitui esta regra pela maximização das diferenças entre os grupos. Sharma (1996) explica que, na análise discriminante, essa solução é obtida pela maximização da soma dos quadrados entre os grupos e pela minimização da soma dos quadrados dentro dos grupos.

Por fim, a análise discriminante pode ser considerada uma técnica confirmatória da análise de conglomerados estudada no Capítulo 6, uma vez que, enquanto esta última baseia-se na definição de grupos de forma exploratória por parte do pesquisador, na análise discriminante esses grupos podem servir de *inputs* para a classificação de novas observações ou para o estudo da significância de variáveis explicativas para sua formação.

11.8. CONSIDERAÇÕES FINAIS

A Análise Discriminante apresenta a vantagem de reduzir a dimensão espacial de análise e elaborar uma sequência de estudos individuais para cada elemento inserido na pesquisa. A técnica apresenta, entretanto, certas desvantagens principalmente em relação a seus pressupostos, uma vez que, em diversos bancos de dados, não há a existência de normalidade multivariada das variáveis explicativas. O pressuposto de existência de homogeneidade das matrizes de covariância também se manifesta, por vezes, como um empecilho para pesquisadores interessados em sua aplicação prática.

Embora as desvantagens provoquem uma maior atenção do pesquisador quando de sua aplicação, a análise discriminante é uma das técnicas mais utilizadas para fins de previsão e classificação de observações em grupos. A partir de suas proposições, a análise discriminante apresenta vasta aplicabilidade em diversas áreas do conhecimento, como biologia, psicologia, educação, finanças, marketing, entre outras, e o desenvolvimento computacional e gráfico de pacotes estatísticos tem, recentemente, propiciado uma crescente usabilidade da técnica nos mais variados problemas de pesquisa.

11.9. EXERCÍCIOS – APLICAÇÃO DE BANCOS DE DADOS

1) O banco de dados do arquivo **EmpresasDiscriminante.sav** apresenta alguns indicadores de 45 empresas, a saber:
   - código da empresa (*Cód_Emp*);
   - prazo médio de recebimento das vendas (*PMRV*, em dias);
   - endividamento (*Endividamento*, em %);
   - vendas (*Vendas*, em R$ x mil);
   - margem líquida das vendas (*Margem_líquida*, em %).

   Esta base de dados é a mesma utilizada no exemplo prático do Capítulo 7 (Análise Fatorial).
   a) Elabore uma análise de *cluster* (método *furthest neighbor* e distância quadrática euclidiana) e salve a posição de cada empresa em uma alternativa com quatro *clusters*.
   b) Na sequência, elabore uma análise discriminante e confirme a análise de *cluster* obtida (percentual na tabela de resultados da classificação).
   c) Verifique os pressupostos da análise discriminante, bem como a significância das variáveis *PMRV*, *Endividamento*, *Vendas* e *Margem_líquida* nas funções discriminantes.
   d) Defina o grupo em que se inseriria uma nova empresa com os seguintes indicadores:
      - PMRV: 29,00 dias;
      - endividamento = 31,00 %;
      - vendas: R$ 2.367,00 (x mil);
      - margem líquida: 10,20 %.
2) O banco de dados do arquivo **BancoDiscriminante.sav** apresenta alguns indicadores de produtividade de três grupos de agências bancárias (A, B e C) de determinado banco, a saber:
   - clientes médios atendidos por hora e por caixa (*Produtividade*);
   - tempo médio na fila por cliente (*Tempo_fila*, em minutos);
   - tempo médio no banco por cliente (*Tempo_banco*, em minutos).

   Para cada grupo de agências, coletaram-se dados referentes a 30 agências. Cada grupo foi definido por meio de uma análise de *cluster* inicial e, portanto, neste exemplo há três grupos predefinidos. Pede-se:
   a) Verifique se há correlação entre as variáveis e, se necessário, aplique análise fatorial anteriormente à aplicação da análise discriminante.
   b) Verifique os pressupostos da análise discriminante, bem como a significância das variáveis *Produtividade*, *Tempo_fila*, *Tempo_banco* nas funções discriminantes.
   c) Analise os resultados da classificação, verificando se os grupos foram classificados corretamente.
   d) Defina o grupo em que se inseriria uma nova agência com os seguintes indicadores:
      - *produtividade*: 25 clientes/hora-caixa;
      - *tempo_fila*: 8 min;
      - *tempo_banco*: 30 min.
3) O banco de dados do arquivo **Faculdades.sav** apresenta alguns indicadores de 50 faculdades localizadas na Região Metropolitana de Fortaleza, a saber:
   - mensalidade (*Mensalidade*, em R$);
   - distância da faculdade até o centro de Fortaleza (*Distância*, em km);
   - total de alunos matriculados (*Alunos*);
   - salário médio por professor tempo integral (*Salário*, em R$).

   Pede-se:
   a) Elabore uma análise fatorial utilizando o método de extração por componentes principais e rotação Varimax e solicite a determinação de dois fatores. Salve os fatores como duas novas variáveis.
   b) Elabore uma análise de *cluster* (método *Between Groups* e distância quadrática euclidiana) para os dois fatores obtidos no item (a) e salve a posição de cada empresa em uma alternativa com três *clusters*.
   c) Na sequência, elabore uma análise discriminante com os dois fatores como variáveis explicativas por meio do procedimento simultâneo e avalie a análise de *cluster* obtida (percentual da tabela de resultados da classificação).

d) Verifique os pressupostos da análise discriminante, bem como a significância dos fatores nas funções discriminantes.
e) Elabore uma análise discriminante com as variáveis originais como explicativas por meio do procedimento *stepwise*. Quais variáveis mostraram-se significantes para definição dos grupos? Houve melhora no percentual geral de classificação quando da elaboração do procedimento *stepwise* com as variáveis originais?

11.10. RESUMO

A análise discriminante (AD) é uma técnica multivariada aplicada quando a variável dependente é qualitativa (não métrica) e as variáveis explicativas são quantitativas (métricas). A AD pode ser considerada uma técnica confirmatória da análise de conglomerados.

O objetivo principal da técnica é identificar as variáveis que discriminam os grupos e, assim, elaborar previsões a respeito de uma nova observação (por exemplo, um produto, uma pessoa ou uma empresa), identificando o grupo mais adequado a que ela deverá pertencer, em função de suas características. Para alcançar esse objetivo, a análise discriminante gera funções discriminantes (combinações lineares das variáveis) que ampliam a discriminação dos grupos descritos pelas variáveis dependentes.

No caso do estudo de apenas dois grupos de variáveis dependentes, a técnica é chamada de Análise Discriminante Simples. Quando houver o interesse na discriminação entre mais de dois grupos, a técnica é denominada de Análise Discriminante Múltipla (MDA – *Multiple-group discriminant analysis*).

A análise discriminante exige dois pressupostos principais: existência de normalidade multivariada das variáveis explicativas e presença de homogeneidade das matrizes de variância e covariância para os grupos, verificada por meio do teste Box's M.

Com relação ao tamanho da amostra, não deve haver uma grande variabilidade de dimensões entre os grupos. Como regra geral, utilizam-se 20 observações para cada variável explicativa e para cada grupo.

Diversos *outputs* gerados a partir de testes e estatísticas devem ser analisados para a identificação das variáveis que discriminam os grupos. São eles: lambda de Wilks, correlação canônica, Qui-quadrado, *eigenvalue*, matriz de correlações, matriz de estrutura e mapa territorial.

Nos *softwares* estatísticos, como o SPSS, existem dois procedimentos para a definição das funções discriminantes: simultâneo e *stepwise*. O procedimento simultâneo considera a inclusão de todas as variáveis explicativas no modelo, independentemente de sua significância. Já o procedimento *stepwise* é utilizado quando o pesquisador deseja avaliar a significância estatística das variáveis por meio da inclusão passo a passo apenas das variáveis significantes.

Finalmente, para avaliar a qualidade do modelo, verifica-se se os grupos originais foram classificados corretamente.

11.11. QUESTÕES COMPLEMENTARES

a) Quais os principais objetivos da análise discriminante?
b) Cite cinco aplicações da análise discriminante.
c) Quais os principais pressupostos da análise discriminante?
d) Quais as principais estatísticas geradas por meio da análise discriminante? Quais os objetivos de cada uma dessas estatísticas?
e) Quais as principais semelhanças e diferenças entre a análise discriminante e a análise de conglomerados?
f) Discorra brevemente sobre as diferenças existentes entre a análise discriminante e a regressão múltipla.