Ana Cristina da Silva Braga

CURVAS ROC: ASPECTOS FUNCIONAIS E APLICAÇÕES

Universidade do Minho

Braga, Dezembro de 2000

Ana Cristina da Silva Braga

CURVAS ROC: ASPECTOS FUNCIONAIS E APLICAÇÕES

Dissertação submetida à Universidade do Minho
para obtenção do grau de doutor
no Ramo de Engenharia de Produção e Sistemas,
Área de Métodos Numéricos e Estatísticos

Universidade do Minho
Braga, Dezembro de 2000

Para chegar à realidade, uma ideia começa por
se apoderar de espíritos fervorosos e escraviza-os;
a partir desse momento, eles pertencem-lhe e não
vêm diante de si se não o objectivo a atingir.
Por vezes, esse objectivo parece intangível: quanto
mais nos adiantamos, mais ele nos parece distante.
Mas que importa?
Os escravos de uma ideia são incapazes de desanimar.

Marie Curie

Esta dissertação é dedicada com todo o carinho

aos meus pais e irmã

ao Carlos e à Catarina

Conteúdo

Lista de Figuras

Lista de Tabelas

Agradecimentos

No decorrer deste trabalho, que começou em 1996, tive oportunidade de poder contar com o apoio de diversas pessoas, que através da confiança em mim depositada e do tempo que me dedicaram, me foram dando força para continuar.

Quero agradecer em especial ao Prof. Pedro Oliveira, meu orientador científico neste trabalho, que me motivou para o desenvolvimento do tema e com o seu profissionalismo e amizade me guiou no decorrer deste.

Agradeço também,

- ao Marco Leal, pela sua colaboração no desenvolvimento do ROCNPA;
- ao Dr. António Gomes da Unidade de Cuidados Intensivos do Hospital Garcia de Orta, de Almada, e à Dr^a Sameiro Carvalho da Unidade de Cuidados Intensivos do Hospital Maria Pia, no Porto, pela cedência de dados e pela sensibilidade que conseguiram transmitir para o tipo de problema que enfrentam;
- a todo pessoal do Grupo disciplinar de Engenharia de Sistemas e Produção Industrial do Departamento de Produção e Sistemas, da Escola de Engenharia, da Universidade do Minho pelo apoio que me deram e pelo seu espírito de camaradagem sempre presente;

- ao Prof. Charles E. Metz da Universidade de Chicago, pela disponibilização de alguns dos seus trabalhos, e dos programas desenvolvidos pela sua equipa;

- ao Carlos e à Catarina, por me aturarem.

Uma última palavra para referir que os trabalhos de investigação apresentados nesta dissertação foram suportados pelo programa PRODEP, concurso n^o 1/96 - PRODEP II.

Resumo

A análise ROC (Receiver Operating Characteristic) é uma ferramenta poderosa para medir e especificar problemas no desempenho do diagnóstico em medicina.

Esta análise por meio de um método gráfico simples e robusto, permite estudar a variação da *sensibilidade* e *especificidade*, para diferentes valores de corte. Neste trabalho é feita a descrição da evolução desta análise, bem como o desenvolvimento do índice *área abaixo da curva ROC*.

A *área abaixo da curva ROC* está associada ao poder discriminante de um teste de diagnostico. Analiticamente, a *área abaixo da curva ROC* pode ser determinada através de:

- métodos de resolução numérica, tipo regra do trapézio;
- métodos estatísticos: relação com a estatística de Wilcoxon-Mann-Witney [37] e estimativa de máxima verosimilhança [26].

Geometricamente, a curva ROC é um gráfico de pares "x" e "y" (que correspondem, a (*1-especificidade*) e à *sensibilidade*, respectivamente) num plano designado por plano ROC unitário. A designação de plano ROC unitário, deve-se ao facto das coordenadas deste gráfico representarem medidas de probabilidade, e por conseguinte variarem entre zero e um.

Uma questão que se colocou no inicio deste trabalho, foi a seguinte:

"Dada a versatilidade e robustez da curva ROC, como poderá ser modelada?"

Para responder a esta questão, procurou-se estudar como é que algumas hipóteses sobre as distribuições associadas à variável de decisão podem afectar a forma da curva ROC.

Com base na hipótese da Normalidade, e através de estudos de simulação, procurou-se numa primeira abordagem verificar qual a variação da forma da curva ROC em função do parâmetro de localização e/ou de escala para a função densidade de probabilidade dos casos designados por anormais (valores maiores na variável de decisão).

Consideraram-se ainda as hipóteses de funções densidade de probabilidade Logísticas e de igual variância, Exponenciais negativas com diferentes parâmetros de escala e Uniformes num intervalo (a, b).

Para a visualização da curva ROC, utilizou-se a representação desta no plano ROC unitário e no plano *binormal*.

São apresentados alguns exemplos ilustrativos, no campo da análise de diagnóstico em medicina, para melhor compreensão da metodologia em estudo. As primeiras aplicações tratam amostras correlacionadas, enquanto que numa outra aplicação é tratado um conjunto de dados independentes.

Após um estudo exaustivo dos programas existentes para a análise ROC, chegou-se à conclusão que poderia ser desenvolvido um novo programa para melhor cumprimento de alguns dos objectivos.

A elaboração de um novo programa, recorrendo a uma nova linguagem de programação (JAVA), permite fazê-lo correr em plataformas diferentes do DOS ou WINDOWS (como por exemplo LINUX, SOLARIS e UNIX). Este programa visa minimizar o trabalho tido para traçar a curva ROC, achar o valor da área abaixo desta pelos diferentes métodos sugeridos no desenrolar

do trabalho e comparar várias curvas ROC em termos do índice área abaixo da curva ROC (para amostras independentes e amostras correlacionadas). Permite ainda, efectuar um ajuste à curva ROC empírica no plano unitário.

Abstract

ROC (Receiver Operating Characteristic) analysis is a powerful tool to measure diagnostic performance in medicine.

This analysis through a robust graphic method, studies the variation of *sensibility* and *specificity*, to different *cut-off* values. In this work the evolution of this analysis is described as well as the relationship with the precision index area under the ROC curve.

The area under the ROC curve is an index of the discriminating power of a diagnostic test. Analytically it can be determined through:

- numerical methods, such as the trapezoidal rule;
- statistical methods such as the Wilcoxon-Mann-Whitney test [37] or the maximal likelihood estimation [26].

Geometrically, the ROC curve is a "x", "y" graphic (representing *1-specificity* and *sensibility*, respectively) in a unitary ROC plane. The unitary designation is due to the fact the coordinates of this graphs are probability measures, and thus its values ranging from zero to one.

A question that this work tries to answer is how can the ROC curve be model. To answer this question it was investigated how some hypothesis concerning the distributions of the decision variable might affect the shape of the ROC curve.

The dependence of the ROC curve shape, from the Normal distribution parameters was studied. Other hypothesis were considered such has Logistic distribution with the same variance, Exponential distribution with different scale parameters and Uniform distribution in a given interval (a, b).

To visualize the ROC curve two representations were used in the unitary ROC plane and as the *binormal* plane.

Some examples, in medical diagnostic are also presented. The first application deals with correlated samples whereas in another application independent samples are studied.

After an exhaustive review of existing software for ROC analysis, and due to the limitations founded, a new software was developed. This new software, based on JAVA, can be run in different platforms from DOS or WINDOWS (for example LINUX, SOLARIS and UNIX). This software allows the drawing of ROC curve, calculate the index area under the ROC curve through different methods, and compares different ROC curves in terms of the index area (for independent and correlated sample). Finally the empirical fit to the curve on the unitary plane is also provided.

Capítulo 1

Introdução

1.1 Motivação

Qualquer investigador deparado com a necessidade da análise de dados, precisa de fazer uma escolha racional sobre o método particular de análise. Devem ser tidas em conta algumas considerações importantes nessa escolha, como por exemplo:

- o objectivo da investigação;
- as características matemáticas das variáveis envolvidas;
- as hipóteses estatísticas feitas sobre estas variáveis;
- como foram recolhidos os dados.

As duas primeiras considerações, são de um modo geral, suficientes para determinar uma análise apropriada. No entanto, o investigador deve também considerar os dois últimos items antes de finalizar a recomendação.

Para certos acontecimentos, existem testes baseados quer em observações de determinado fenómeno, quer em técnicas laboratoriais, que permitem a

previsão ou detecção desse acontecimento numa fase incipiente de desenvolvimento. São exemplos, alguns testes epidemiológicos que são a base de rastreio para o diagnóstico precoce de algumas doenças.

Uma questão problemática, e que funcionou como um estímulo para o desenvolvimento deste estudo, é o problema da discriminação existente num teste de diagnóstico, que consiste em conseguir classificar de uma forma precisa os casos considerados normais e os anormais.

Outra questão que se torna problemática num teste de diagnóstico, prende-se com as definições de *exactidão* e *precisão*. A *precisão* está associada à dispersão dos valores em sucessivas observações, enquanto que a *exactidão* refere-se à proximidade de uma estimativa do verdadeiro valor que pretende representar. As limitações da *exactidão* e da *precisão* no diagnóstico, originaram a introdução dos conceitos de *sensibilidade* e *especificidade* dum teste de diagnóstico. Estas medidas e os índices a elas associados, como a *proporção de verdadeiros positivos* e a *proporção de falsos positivos*, são mais significantes do que a *exactidão*, embora não forneçam uma descrição única do desempenho de diagnóstico.

O maior problema da *sensibilidade* e da *especificidade* é que estas medidas dependem do critério de diagnóstico ou de um *valor de corte*, o qual é por vezes seleccionado arbitrariamente. Assim, mudando o critério pode-se aumentar a *sensibilidade* com o consequente detrimento da *especificidade*, e vice-versa. Consequentemente, estas medidas representam um quadro incompleto do desempenho de um teste de diagnóstico.

Deverá ainda ter-se em consideração, que um critério de decisão particular depende também dos benefícios associados aos resultados correctos e dos custos associados aos incorrectos. Por exemplo, a previsão de uma tempestade que acaba por não ocorrer (*falso positivo*) é tipicamente vista como tendo

um custo menor do que em relação à falha na previsão de uma tempestade que ocorre (*falso negativo*), assim o critério a adoptar para um diagnóstico positivo deverá estar do lado mais brando.

Num teste de diagnóstico existem dois tipos de erro que podem ocorrer na decisão, a escolha de uma *falha* (no sentido de declarar um doente como são) ou a escolha de um *falso alarme* (declarar uma pessoa sã como doente). Por exemplo, para um profissional que tem perante si um dado diagnóstico para uma doença, ao ter de decidir, ele irá preferir um *falso alarme* a uma *falha* - principalmente se a doença for contagiosa - pois este tipo de erro conduzirá, para este profissional, ao que se poderá designar por "*um mal menor*" em termos de diagnóstico. Isto é, ele irá optar certamente por um teste mais sensível. Por outro lado, ele deverá estar consciente que uma terapia disponível para este tipo de doença poderá ser efectivamente, cara e deficiente, o que torna o teste pouco específico.

Para contornar este tipo de situações, foi necessário desenvolver medidas alternativas de diagnóstico com propriedades mais robustas do que a *sensibilidade* e a *especificidade* per si. A análise ROC (*Receiver Operating Characteristic*) foi a técnica desenvolvida para tornear este tipo de problema.

A análise ROC pode ser efectuada através de um método gráfico simples, e o desempenho de um dado teste poderá ser avaliado através de índices de precisão simples associados à curva ROC, como por exemplo a área abaixo desta.

Considere-se como exemplo um trabalho desenvolvido por Ribeiro et al (1993) [73]. Neste trabalho é apresentado o estudo de uma doença rara GM2 - gangliosidosis com a variante $B1$, que geralmente se encontra associada a um determinado grupo étnico-geográfico, e que parece ser excepcionalmente frequente em Portugal. Com o objectivo de estabelecer um método de de-

tecção desta doença, os autores aplicaram um teste designado por Hex A para identificação dos portadores nas famílias com variante $B1$, a um grupo designado por "$B1$ variant carrier" e a um outro grupo de controle, e compararam estes resultados com os obtidos através da análise de DNA.

Chegaram a resultados de *sensibilidade* $(0,996)$ e *especificidade* $(0,994)$ do teste para um *valor de corte* específico $(0,195)$. A fiabilidade aqui obtida, do teste Hex A para identificação dos portadores nas famílias com variante $B1$, e a sua utilização devido aos baixos custos e a atractiva possibilidade de automatização, levaram os investigadores a colocarem a hipótese de este ser uma alternativa ao usual teste de DNA que envolve maiores custos e morosidade no processo.

Para responder à questão, se este pode ser um teste fiável, isto é, com elevado poder de discriminação para identificação deste tipo de doença, as medidas de *sensibilidade* e *especificidade,* determinadas pelos autores, num *valor de corte* específico não se tornam suficientes. No entanto, se fosse efectuada uma análise estatística baseada numa curva ROC para teste Hex A, por exemplo por determinação do índice área abaixo da curva, poder-se-ia avaliar de uma forma mais precisa o desempenho deste como teste de diagnóstico alternativo ao teste de DNA.

Considere-se o problema, formulando as seguintes hipóteses:

H_0 : O indivíduo apresenta a doença, D
H_1 : O indivíduo não apresenta a doença, $\overline{D}$

Assim, para um *valor de corte* específico a representação ROC dá a probabilidade de não rejeitar H_0, isto é, considerar que o indivíduo apresenta a doença.

De uma forma geral, o grande motivo que levou ao desenvolvimento deste trabalho de doutoramento foi a necessidade de explorar e sistematizar a análise da curva ROC e os índices a ela associados, dado o vasto campo de aplicabilidade desta análise e a facilidade de tratamento matemático, procurando justificar a sua robustez. Por outro lado, a necessidade de encontrar um programa que possibilite a sistematização dos cálculos, assim como, a apresentação dos resultados gráficos da análise ROC, levou ao desenvolvimento de um programa de apoio para esta análise.

1.2 Objectivos

Tendo em conta o plano de doutoramento inicialmente traçado, delinearam-se os seguintes objectivos:

1. Explicitar a relação entre a estatística U de Wilcoxon-Mann-Whitney e o valor da área abaixo da curva ROC.

2. Procurar uma expressão analítica para a curva ROC que traduza uma relação entre a *sensibilidade*e a *especificidade* no plano ROC unitário, e analisar se na realidade a sua forma varia consoante a distribuição associada aos dados.

3. Procurar uma *"curva de ajuste"* à curva ROC empírica.

4. Tratar algumas aplicações através da metodologia ROC, e analisar os resultados obtidos.

5. Desenvolvimento de um programa com implementação em diferentes plataformas (WINDOWS, LINUX, UNIX e Macintosh) para o estudo da curva ROC e comparações de testes através desta metodologia. O

programa procurará minimizar o trabalho necessário para o desenho da curva ROC, calcular o valor da área abaixo da curva ROC pelos diferentes métodos sugeridos, e comparar várias curvas ROC (para amostras independentes e amostras correlacionadas).

1.3 Estrutura da dissertação

Esta dissertação desenvolve-se ao longo de oito capítulos. O conjunto de objectivos propostos na secção anterior traduzem, ainda que parcialmente, o modo como o trabalho foi estruturado. Nesta secção ao apresentar a organização da dissertação, pretende-se orientar o leitor nas linhas seguidas ao longo do seu desenvolvimento.

Assim, após esta introdução, o segundo capítulo apresenta uma perspectiva histórica sobre a análise ROC assim como a sua relação com a teoria estatística, a teoria de detecção do sinal e a análise de diagnóstico.

No capítulo 3 é apresentado uma breve descrição do estado da arte. São referidos os trabalhos resultantes de uma longa pesquisa bibliográfica sobre o tema análise ROC.

No quarto capítulo serão apresentados os principais contributos para o desenvolvimento da análise ROC. São analisados trabalhos como o de Bamber [8], Green e Swets [33], Metz [58], Iverson [47] e DeLong [22].

O capítulo 5 surge como resposta ao segundo e terceiro pontos referidos nos objectivos. Procura-se determinar uma relação entre as funções densidade de probabilidade associadas aos dados e a forma da curva ROC e, por outro lado, determinar um *"ajuste"* à curva ROC empírica.

No capítulo 6 são apresentadas algumas aplicações recorrendo à análise de dados através da curva ROC, nomeadamente do índice de precisão área

abaixo da curva. São utilizados conjuntos de dados correlacionados e independentes.

O capítulo 7 faz a apresentação dos programas estudados para o desenrolar do trabalho proposto, assim como as suas principais características.

Para responder às dificuldades tidas no capítulo 6, no que diz respeito à análise de dados ROC, foi desenvolvido um programa que se designou por ROCNPA, e que é apresentado no capítulo 8. Neste capítulo são analisadas diferentes situações utilizando os programas já existentes e o ROCNPA. São confrontados os resultados obtidos e apresentadas as vantagens e desvantagens do novo programa.

Por fim, são apresentadas algumas conclusões gerais sobre o trabalho realizado, e apontam-se algumas linhas de orientação para futuros trabalhos de investigação.

Capítulo 2

Teoria da Análise ROC

A análise ROC (*Receiver Operating Characteristic*) teve origem na teoria de decisão estatística e foi desenvolvida entre 1950 e 1960 para avaliar a detecção de sinais em radar e na psicologia sensorial [58]. A potencial utilidade da análise ROC em avaliar diagnósticos médicos foi desde então utilizada por vários autores [58] e, subsequentemente, foi aplicada com sucesso a uma grande variedade de testes de diagnóstico [82] e em particular no diagnóstico de imagem médica [67], [11].

Consequentemente, as várias aplicações da análise ROC à medicina estimularam o desenvolvimento de metodologias de análise estatística dos dados ROC [58], [37].

2.1 Perspectiva Histórica

A necessidade da obtenção de medidas de discriminação precisas é já um problema de longa data. Um dos campos em que esta medida foi primeiramente desenvolvida foi no campo da psicologia sensorial.

Gustav Theodor Fechner (1801-1887) foi considerado o pioneiro. O seu

objectivo era derrubar o materialismo e concebeu a psicofísica para o ajudar, procurando demonstrar uma relação empírica entre a mente e o corpo [80].

Para Fechner, um estímulo deveria ser apresentado a cada sujeito centenas de vezes, com o objectivo de obter uma estimativa relativamente estável de respostas positivas. Considerou como respostas positivas afirmações do tipo:

"Sim, reconheço o estímulo A (como oposto ao B)"

"O estímulo A é maior do que o estímulo B".

Posteriormente este procedimento foi designado por critério de *comparação múltipla* ou *escolha-forçada* [80].

Fechner representou graficamente a proporção de respostas positivas contra a medida física da intensidade do estímulo e obteve assim, uma função psicométrica [80].

No seguimento dos trabalhos de Fechner, Louis Leon Thurstone (1887-1955), continuou o estudo da análise discriminante. Demonstrou como os métodos de Fechner poderiam ser utilizados para quantificar os atributos psicológicos do estímulo na realidade não susceptível a medição física, por exemplo, para determinar a superioridade da escrita com a mão direita [80].

O modelo de Thurstone começa por colocar a hipótese de sobreposição das distribuições de intensidade psicológica de dois estímulos semelhantes, como exemplificado na figura 2.1. O modelo prossegue com algumas hipóteses específicas, incluindo a normalidade, não existência de correlação entre os estímulos e igualdade de variâncias.

No modelo representado na figura 2.1, assume-se que o efeito sensorial varia de acordo com a distribuição da esquerda quando o estímulo nulo, S_0, ou *"ruído"*, está presente, e varia de acordo com a distribuição da direita, S_1, quando um dado "*sinal*" é adicionado ao *"ruído"*. O critério para uma

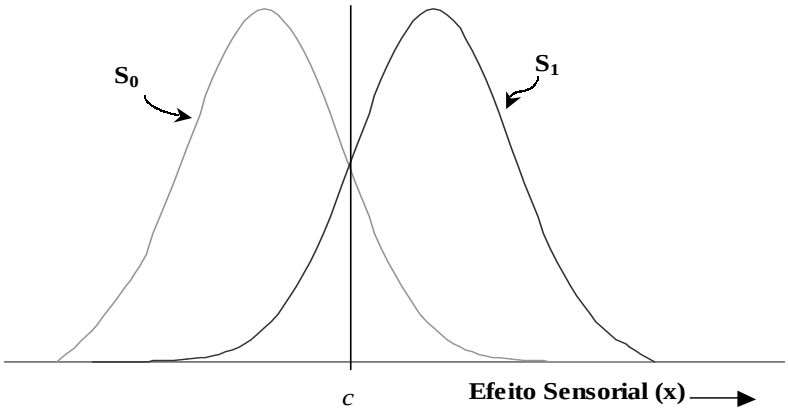


Figura 2.1: Um exemplo do modelo de reconhecimento de Thurstone para representar a detecção.

resposta positiva, c, é considerado fixo em determinado ponto onde raramente é excedido pelo *"ruído"*, sem discriminação possível abaixo deste ponto [80].

A hipótese de simetria de Thurstone, é equivalente a assumir um *critério de decisão*, isto é, a considerar um *valor de corte, c*, que corresponde ao ponto onde as duas distribuições se cruzam. [80].

O passo seguinte em psicofísica foi dado em 1940 por H. Richard Blackwell. Blackwell defendeu um procedimento semelhante ao da comparação-múltipla, que designou por *escolha-forçada.* Debruçou-se sobre o problema de detecção, no qual um dos dois estímulos é considerado como sendo o *estímulo nulo.*

As duas distribuições, na abordagem de Blackwell, não têm de ser obrigatoriamente iguais; uma, designada por *"ruído",* poderia ser Normal como no modelo de Thurstone enquanto que a outra com maior média, eventualmente com variância diferente, poderia representar o que designou por *"sinal"* [80].

A partir de 1950, os estudos no campo da psicofísica nomeadamente no domínio da audição e visão, foram conduzidos por vários autores, entre eles

John Swets [80].

2.2 Teoria Estatística

O problema em termos de testes de hipóteses, ou tomada de decisões estatísticas, pode ser representado da mesma forma que Thurstone e Blackwell representam o problema de discriminação. A figura 2.2 mostra uma representação hipotética para este tipo de problema. A distribuição da esquerda representa nesta situação, a hipótese nula, H_0, e a da direita uma hipótese alternativa, H_1.

Assim, as hipóteses do problema poderão ser especificadas como:

H_0: A população tem média $\mu = \mu_0$;

H_1: A população tem média $\mu = \mu_1$.

Com base numa observação x, uma das hipóteses é aceite.

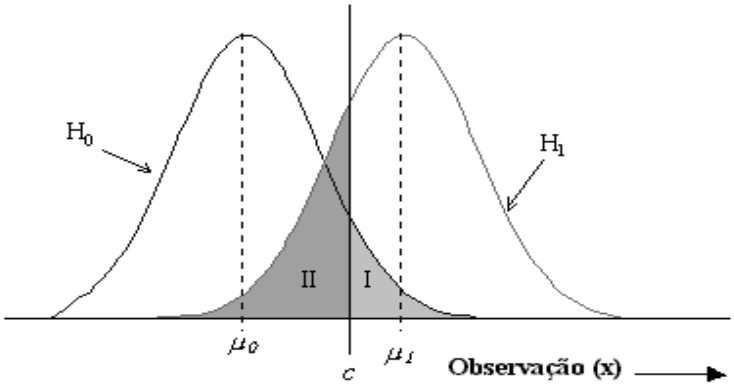


Figura 2.2: Distribuições de duas populações.

Como pode ser verificado na figura 2.2, H_0 é a hipótese nula que considera que a população tem média $\mu = \mu_0$ e H_1 é a hipótese alternativa

que considera que a população tem média $\mu = \mu_1$. Assim, a área sombreada à direita do critério de decisão, c, representa a probabilidade de cometer um *erro de tipo I*, que corresponde à probabilidade de rejeitar H_0 quando H_0 é verdadeira; a área sombreada à esquerda do critério de decisão, c, representa a probabilidade de cometer um *erro de tipo II*, que corresponde à probabilidade de não rejeitar H_0 quando H_1 é verdadeira.

A construção do teste estatístico é equivalente a dividir o eixo x em duas regiões, separadas pelo critério de decisão c. Valores de x menores que c conduzirão à aceitação da hipótese nula, H_0, e valores de x maiores que c conduzirão à aceitação da hipótese alternativa, H_1. Consoante o critério de decisão escolhido, pode-se determinar a probabilidade de cometer um *erro de tipo I* ou *tipo II* (figura 2.2).

Existem princípios gerais para os testes de hipóteses que obedecem a determinadas regras desenvolvidas por Neyman e Pearson. A principal regra associada a estes, e a mais familiar em estatística, é fixar a probabilidade de cometer um *erro de tipo I* arbitrariamente (a um nível de significância usualmente de 0.05 ou 0.01) e depois escolher um critério de forma a minimizar a probabilidade de cometer um *erro de tipo II*. Estes autores demonstraram que o melhor teste é definido em termos da razão da verosimilhança. Aceita-se H_1 quando a razão das verosimilhanças excede determinado valor c, que é escolhido para produzir a probabilidade desejada de cometer um *erro de tipo I* [80].

A *potência do teste* é definida por:

$$k = \begin{cases} \text{Prob } (\textit{erro de tipo I}) & \text{sob } H_0 \\ 1 - \text{Prob } (\textit{erro de tipo II}) & \text{sob } H_1 \end{cases}$$

Sob as regras de Neyman-Pearson, fixa-se a probabilidade de cometer um *erro de tipo I* e escolhe-se a razão de verosimilhança igual a c de forma a

maximizar a *potência do teste*.

Assim é possível definir a *curva característica de operação,* que não é mais do que a representação gráfica do complementar da *função potência do teste* $(1-k)$. A curva ROC é uma maneira gráfica de comparar duas curvas características de operação - a que se definiu anteriormente, em que se fixa a probabilidade de cometer um *erro de tipo I* arbitrariamente, e uma outra que mostra a variação em probabilidade de um *erro de tipo I* para um valor fixo de probabilidade de cometer um *erro de tipo II.* Na figura 2.3*,* encontram-se representadas as *curvas características de operação* para as duas situações descritas, considerando um teste hipotético para duas distribuições Normais com igual variância.

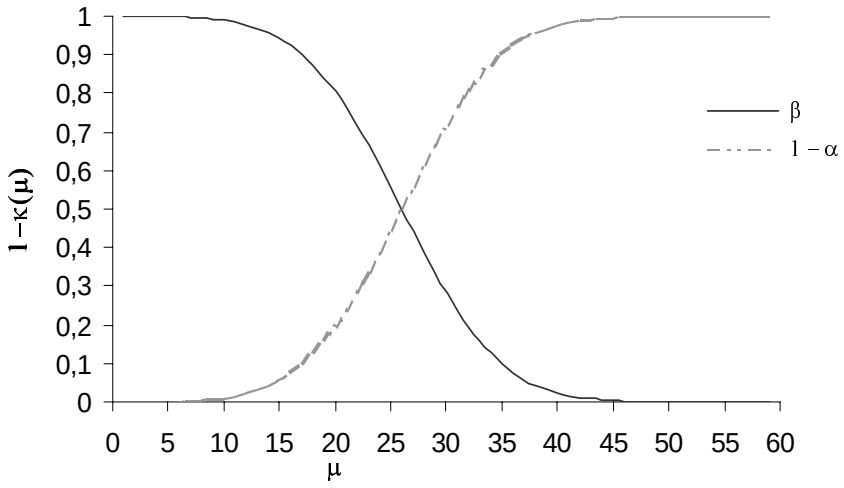


Figura 2.3: Representação de curvas características de operação.

A curva ROC, transmite a informação da conjunção destas duas curvas de operação, isto é, mostra como podem variar os dois tipos de erro, com a mudança de critério de decisão.

Um outro avanço na teoria da decisão estatística, por volta de 1940, foi

dado por Abraham Wald. Wald demonstrou que algumas regras de decisão diferentes - como a maximização da proporção de decisões correctas, maximização do valor esperado de uma decisão e maximização da mínima recompensa - são unificadas pela razão das verosimilhanças [80]. Posteriormente, Green e Sweets [33], descrevem algumas das regras de decisão mais utilizadas em estatística que a seguir se apresentam.

1. *Maximização de uma combinação ponderada.*

No caso da existência de duas alternativas, os resultados poderão ser descritos por quatro probabilidades diferentes. Apenas duas dessas probabilidades são independentes, dado que:

$$P(H_0 \mid h_0) + P(H_1 \mid h_0) = 1$$

e

$$P(H_0 \mid h_1) + P(H_1 \mid h_1) = 1$$

Assim, o objectivo seria, sempre que possível, maximizar $P(H_1 \mid h_1)$, ao mesmo tempo que se minimizaria $P(H_1 \mid h_0)$. Geralmente, não se consegue satisfazer os dois objectivos simultaneamente, pelo que se opta pela maximização da quantidade:

$$\{P(H_1 \mid h_1) - \gamma \; P(H_1 \mid h_0)\} \tag{2.1}$$

onde γ é uma constante, $\gamma > 0$ [33].

Designando por A o conjunto de todos acontecimentos que conduzem à aceitação de h_1, então a probabilidade de H_1 ser aceite quando h_1 é verdadeira é dada por,

$$\sum_{e_i \in A} P(e_i \mid h_1) = P(H_1 \mid h_1) \qquad \text{(para o caso discreto)}$$
$$\int_{e_i \in A} P(e_i \mid h_1) = P(H_1 \mid h_1) \qquad \text{(para o caso contínuo).}$$

De forma análoga, a probabilidade de uma aceitação incorrecta da hipótese h_1, é dada por,

$$\sum_{e_i \in A} P(e_i \mid h_0) = P(H_1 \mid h_0) \qquad \text{(para o caso discreto)}$$
$$\int_{e_i \in A} P(e_i \mid h_0) = P(H_1 \mid h_0) \qquad \text{(para o caso contínuo).}$$

Com o objectivo de maximizar $P(H_1 \mid h_1)$, deve-se escolher a região A de forma que:

$$P(H_1 \mid h_1) - \gamma\ P(H_1 \mid h_0) = \sum_{e_i \in A} P(e_i \mid h_1) - \gamma \sum_{e_i \in A} P(e_i \mid h_0)$$
$$\text{(para o caso discreto)}$$
$$= \int_{e_i \in A} P(e_i \mid h_1) - \gamma \int_{e_i \in A} P(e_i \mid h_0)$$
$$\text{(para o caso contínuo)}$$

seja tão grande quanto possível.

Note-se que apenas se deve incluir em A, acontecimentos cuja razão de verosimilhanças de um acontecimento e_k para a hipótese h_1 em relação à hipótese h_0 - $l_{10}(e_k)$ - satisfaçam a condição[33]:

$$l_{10}(e_k) = \frac{P(e_k \mid h_1)}{P(e_k \mid h_0)} \geq \gamma.$$

Assim, a primeira regra de decisão pode ser definida da seguinte forma:

Uma regra de decisão que maximize $P(H_1 \mid h_1) - \gamma\, P(H_1 \mid h_0)$*, consiste em escolher* H_1 *se e só se a razão de verosimilhanças para todos acontecimentos* e_i*,* $l_{10}(e_i) \geq \gamma$*,* onde γ é o valor do critério adoptado.

2. *Maximização do valor esperado.*

Considere-se uma situação de decisão binária para a qual certos valores e custos estão definidos para os quatro resultados possíveis. Na notação que se apresenta o primeiro subscrito corresponde à alternativa apresentada e o segundo à alternativa escolhida [33].

V_{00} valor associado à escolha correcta de H_0;

V_{01} valor (custo) associado à escolha incorrecta de H_1 (quando de facto H_0, é a alternativa correcta); significa que a pessoa perde V_{01} quando este tipo de escolha incorrecta é efectuada.

V_{11} valor associado à escolha correcta de H_1;

V_{10} valor (custo) associado à escolha incorrecta de H_0 (quando de facto H_1, é a alternativa correcta); significa que a pessoa perde V_{10} quando este tipo de escolha incorrecta é efectuada.

O valor esperado de uma estratégia de decisão, (d), é definido por:

$$\begin{aligned} E\,(d) = {} & V_{00}\ P(h_0)\ P(H_0 \mid h_0) + V_{11}\ P(h_1)\ P(H_1 \mid h_1) \\ & - V_{10}\ P(h_1)\ P(H_0 \mid h_1) - V_{01}\ P(h_0)\ P(H_1 \mid h_0) \end{aligned} \tag{2.2}$$

Supondo que o objectivo é maximizar o valor esperado dado pela expressão da equação (2.2), então a questão que se levanta é como determinar

as regiões de aceitação de h_0 e h_1 de forma a atingir esse objectivo. Note-se que maximizar o valor esperado, considerando os valores dos custos e as probabilidades *a priori* fixas, é equivalente a maximizar uma expressão do tipo da definida na equação (2.1) [33].

Para demonstrar esta equivalência, consideram-se as relações

$$P(H_0 \mid h_0) = 1 - P(H_1 \mid h_0)$$

e

$$P(H_0 \mid h_1) = 1 - P(H_1 \mid h_1)$$

donde resulta, que:

$$\begin{aligned} E\,(d) = {} & V_{00}\ P(h_0) - V_{00}\ P(h_0)\ P(H_1 \mid h_0) + V_{11}\ P(h_1)\ P(H_1 \mid h_1) \\ & - V_{10}\ P(h_1) + V_{10}\ P(h_1)\ P(H_1 \mid h_1) - V_{01}\ P(h_0)\ P(H_1 \mid h_0). \end{aligned}$$

Como $V_{00}\ P(h_0)$ e $V_{10}\ P(h_1)$ são constantes, maximizar o valor esperado é equivalente a maximizar a quantidade:

$$\begin{aligned} & [V_{11}\ P(h_1) + V_{10}\ P(h_1)]\ \ P(H_1 \mid h_1) \\ & - [V_{00}\ P(h_0) + V_{01}\ P(h_0)]\ \ P(H_1 \mid h_0) \end{aligned} \tag{2.3}$$

Rearranjando, virá:

$$P(H_1 \mid h_1) - \frac{(V_{00} + V_{01})\ \ P(h_0)}{(V_{11} + V_{10})\ \ P(h_1)}\ P(H_1 \mid h_0).$$

Maximizar a equação (2.3), é equivalente a maximizar uma expressão da forma dada na equação (2.1) com

$$\gamma = \frac{(V_{00} + V_{01})\ P(h_0)}{(V_{11} + V_{10})\ P(h_1)}. \tag{2.4}$$

Consequentemente, o valor esperado é maximizado pela aceitação de h_1 para todos os acontecimentos cuja razão de verosimilhanças de h_1 em relação a h_0 é igual ou superior ao valor de γ, como definido na equação (2.4) [33].

3. *Maximização da percentagem de respostas correctas.*

Considerando que os custos associados aos erros são nulos e o valor de uma decisão correcta igual a um, maximizar o valor esperado de uma estratégia de decisão é equivalente a maximizar a percentagem de respostas correctas. Atendendo a que nestas condições,

$$\gamma = \frac{P(h_0)}{P(h_1)}$$

se $P(h_1)$ aumentar, é necessário uma menor razão de verosimilhanças para que H_1 seja escolhido.

2.3 Teoria de Detecção de Sinal

A detecção de sinais electromagnéticos na presença de um *ruído* foi analisada, em 1940 como um problema de teste de hipóteses estatísticas. O *ruído* foi identificado como sendo a hipótese nula, H_0, enquanto o *ruído mais sinal* estava associado com a hipótese alternativa, H_1.

Por exemplo, no contexto dos radares, os *erros de tipo I* são designados *"falsos alarmes",* enquanto que os *erros de tipo II* são *"falhas"*, e ambos são considerados perigosos numa situação de defesa, dado que os seus custos variam com os diferentes tratamentos e as reacções disponíveis ao tratamento.

Na teoria de detecção do sinal, o observador tem como tarefa, decidir com base na aleatoriedade, qual dos estímulos é resultado do *ruído mais sinal*, ou do *ruído*. O problema fundamental de detecção, pode ser visto da seguinte forma [28]:

- Existe uma ocorrência aleatória de dois acontecimentos, *ruído mais sinal (sn)* e *ruído (n)*, e cada acontecimento ocorre num intervalo de tempo bem definido;
- O estímulo físico, ou *evidência* relativo a cada acontecimento, varia de experiência para experiência, e tem um resultado, que é a representação probabilística do acontecimento;
- Após cada observação, o observador deve tomar uma *decisão* do tipo *"sim"* ou *"não"*.

Assim, o procedimento de decisão, envolve dois elementos básicos: *acontecimento* $\longrightarrow$ *decisão*. Cada estímulo deve ser classificado em uma de duas categorias, *sn* ou *n*.

Designando por $P(sn)$, a probabilidade associada à presença de sinal, e $P(n)$, a probabilidade associada à ausência de sinal (só *ruído*), no caso de dois acontecimentos:

$$P(sn) + P(n) = 1.$$

Estas probabilidades são usualmente dadas pela experiência ou natureza e, normalmente, não se encontram sob controle do observador.

Um modelo do tipo *acontecimento* $\longrightarrow$ *decisão*, poderá ser descrito em termos de árvore de probabilidades como ilustra a figura 2.4.

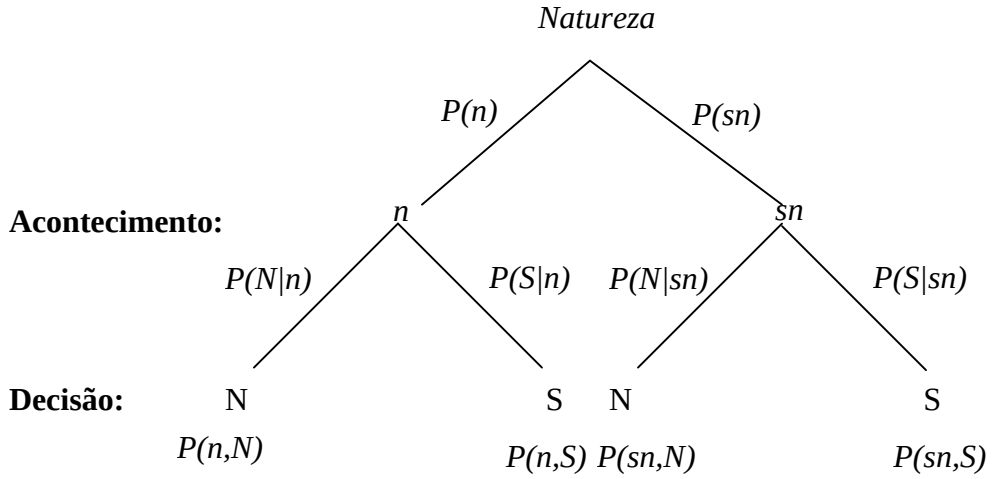


Figura 2.4: Árvore de probabilidades que descreve o comportamento de um observador, no procedimento fundamental de detecção.

Nesta situação a *decisão* do observador é do tipo: *"sim, o sinal encontra-se presente", S* ou *"não, o sinal encontra-se ausente", N.*

O desempenho de um observador numa experiência num único intervalo é usualmente medido em termos de probabilidades conjuntas de *acontecimento−resposta*. Estas probabilidades são baseadas quer no valor da probabilidade *a prior* da existência de sinal, $P(sn)$, quer nos valores das duas probabilidades condicionadas, $P(S \mid sn)$ e $P(S \mid n)$. Assim, define-se:

- *aceitação correcta:* $P(sn, S) = P(S \mid sn)\ P(sn)$;
- *rejeição incorrecta:* $P(sn, N) = [1 - P(S \mid sn)]\ P(sn)$;
- *rejeição correcta:* $P(n, N) = [1 - P(S \mid n)]\ P(n)$;
- *aceitação incorrecta:* $P(n, S) = P(S \mid n)\ P(n)$.

Note-se que, por exemplo, $P(sn, S)$ designa a probabilidade de *acertar*, enquanto que $P(S \mid sn)$ representa a *fracção de acertos.*

No problema fundamental de detecção com dois acontecimentos e duas respostas, definido anteriormente, existem quatro resultados possíveis. Esta situação pode ser descrita na forma de uma tabela de contingência 2×2, padrão onde os dois acontecimentos (*sn* e *n*) possíveis alternativos representam as colunas e as duas respostas permitidas (*sim e não*) representam as linhas, como se ilustra na tabela 2.1.

Tabela 2.1: Tabela de contingência 2x2 na teoria de detecção de sinal

		Acontecimento	
		Ruído+Sinal (sn)	Ruído (n)
Resposta	Sim (S)	$a = P(S \mid sn)$ (acerto ou verdadeiro positivo)	$b = P(S \mid n)$ (falso alarme ou falso positivo)
	Não (N)	$c = P(N \mid sn)$ (valor omisso ou falso negativo)	$d = P(N \mid n)$ (verdadeiro negativo)

Seja X a designação para uma variável aleatória. Se X for discreta, então $P(x \mid sn)$ é a probabilidade condicional de x dado o acontecimento sn; $P(x \mid n)$ é a probabilidade condicional para o mesmo valor de x, dado o acontecimento n. Se a variável aleatória for contínua, então os correspondentes elementos de probabilidade são $f(x \mid sn)\ dx$ e $f(x \mid n)\ dx$, onde $f(x \mid sn)$ e $f(x \mid n)$ são as funções densidade de probabilidade associadas, respectivamente, às duas distribuições da variável aleatória X.

Na situação em estudo interessa que as duas distribuições da variável aleatória X apresentem uma área de sobreposição. Um valor de x neste intervalo de sobreposição deverá transportar a informação na base da qual o observador deve efectuar uma decisão racional.

Define-se a *função razão de verosimilhança*, por:

$$l(x) = \frac{P(x \mid sn)}{P(x \mid n)} \qquad \text{para o caso discreto} \tag{2.5}$$

e

$$l(x) = \frac{f(x \mid sn)}{f(x \mid n)} \qquad \text{para o caso contínuo.} \tag{2.6}$$

A *razão de verosimilhança* $l(x)$ é uma função do valor numérico de x, e esta função exprime a mudança de razão entre as correspondentes funções de probabilidade, ou funções densidade de probabilidade, das duas distribuições de X.

2.4 Análise ROC

A análise *ROC* (*Receiver Operating Characteristic)*, teve a sua origem na teoria de detecção de sinal. Assim, a ROC pode provir de uma tabela de contingência 2×2, do tipo da ilustrada na tabela 2.1.

A *ROC* é assim baseada em duas quantidades que contêm toda a informação da tabela 2.1, uma designada por *fracção de verdadeiros positivos (FVP)*, definida por $a/(a+c)$, e outra designada por *fracção de falsos positivos (FFP)*, definida por $b/(b+d)$, a *fracção de falsos negativos* e a *fracção de verdadeiros negativos* são os respectivos complementares.

Pode-se definir a *ROC* (*Receiver Operating Characteristic)* de duas formas diferentes, uma mais restritiva, em termos da *razão de verosimilhanças*, e uma outra mais geral, em termos da variável de decisão x [28].

Definição 1 *Definição de ROC em termos de* $l(x)$ *- Uma ROC sumaria o conjunto possível de matrizes* 2×2*, que resulta quando um valor de corte* $c =$

$l(x_0)$ varia de uma forma contínua do seu maior valor possível até ao menor possível. Este conjunto de matrizes 2×2 é único para as duas distribuições de X.

Definição 2 *Definição de ROC em termos de x - Uma ROC sumaria o conjunto possível de matrizes 2×2, que resulta quando intervalos disjuntos do eixo do x são sucessivamente adicionados ao intervalo de aceitação; a inclusão de intervalos começa com o intervalo vazio e termina com todo o eixo do x. Os conjuntos possíveis de matrizes 2×2 estão restringidos pelas duas distribuições de X.*

Por exemplo, dado um par de distribuições de X contínuas, apenas uma ROC resulta da utilização de $l(x)$ como critério de decisão. Dado o mesmo par de distribuições, existe um grande número de ROC $'$s, cada uma dependendo da ordem de inclusão dos intervalos em x no critério de aceitação [28].

O sistema de coordenadas da ROC apresenta como ordenadas a proporção de *acertos*, $P(S \mid sn)$, e como abcissas a proporção de *falsos alarmes* $P(S \mid n)$. Quando as probabilidades são projectadas linearmente, os valores de coordenadas variam de zero até um, e todas as ROC possíveis estão limitadas por um quadrado unitário. A diagonal positiva deste quadrado é a designada linha do acaso, em que $P(S \mid sn) = P(S \mid n)$; a diagonal negativa, corresponde a $P(S \mid sn) = 1 - P(S \mid n)$.

A figura 2.5, ilustra o sistema de coordenadas utilizado para representar uma ROC. Cada ponto neste espaço ROC corresponde a uma matriz 2×2. Se o observador utilizar uma regra de decisão *pura,* isto é, se o observador for coerente nas suas respostas para cada x, então, de acordo com as definições dadas de ROC, esta deverá começar em $(0, 0)$ e terminar em $(1, 1)$; sob estas condições a ROC deverá ser não decrescente em todo o seu percurso [28].

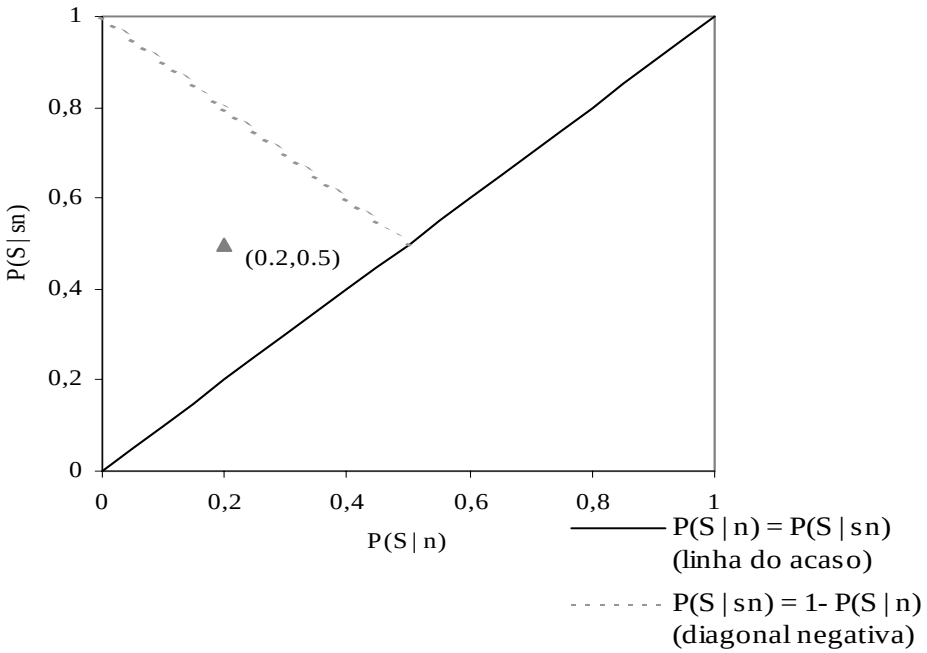


Figura 2.5: Sistema de coordenadas num quadrado unitário, utilizado para representação de uma ROC.

Tabela 2.2: Tabela de contingência 2×2 correspondente a um ponto no espaço ROC.

	Ruído+Sinal (sn)	Ruído (n)
Sim (S)	0,5	0,2
Não (N)	0,5	0,8

Se considerar por exemplo, a matriz 2×2 para uma determinada regra de decisão, dada pela tabela 2.2, o ponto que lhe corresponde no espaço ROC apresenta de coordenadas $(0,2; 0,5)$ como ilustrado na figura 2.5.

A ROC para um observador, em termos da razão de verosimilhanças, su-

maria uma relação especifica entre as duas distribuições de probabilidade. Dado que a análise ROC apresenta muitas aplicações nos mais variados domínios, foi proposto que a designação ROC significava *Relative Operating Characteristic* [28].

2.5 Análise de diagnóstico

2.5.1 Problema em estudo

Considere-se a variável em estudo representada por x e que valores baixos de x favorecem a decisão "normal" (T^-) e valores elevados de x favorecem a decisão "anormal" (T^+).

Designe-se ainda, por $f\ (x|A)$ a distribuição dos valores de x para os casos designados anormais, x_A, e por $f\ (x|N)$ a distribuição dos valores de x para os casos designados normais, x_N; ou seja, a distribuição de x_A deverá ser centrada à direita da de x_N .

Graficamente, a situação descrita, poderia ser ilustrada pela figura 2.6.

Como se pode verificar a partir desta figura, as distribuições de x_A e x_N, sobrepõem-se, e isto significa que, alguns dos casos inicialmente identificados como normais poderão ter leituras como anormais, e por outro lado, alguns dos casos inicialmente identificados como anormais poderão ter leituras como normais.

Para qualquer teste de diagnóstico é fixado um valor de corte para a variável em estudo, valor que determina a classificação dos indivíduos como anormais ou normais. Assim, qualquer teste é avaliado pela comparação relativa da *fracção de verdadeiros positivos (FVP), fracção de falsos positivos (FFP), fracção de verdadeiros negativos (FVN)* e *fracção de falsos negativos (FFN).*

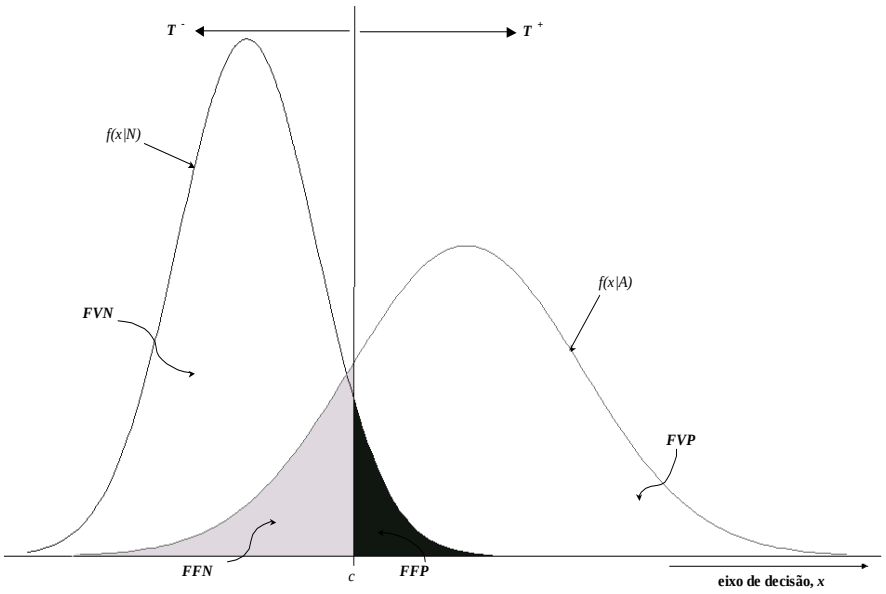


Figura 2.6: Sobreposição de duas distribuições hipotéticas.

Em termos de diagnóstico, a *fracção de verdadeiros positivos (FVP)* corresponde à probabilidade de decidir que a característica em questão está presente, quando de facto está presente. Por outro lado, a *fracção de verdadeiros negativos (FVN)* corresponde à probabilidade de decidir que a característica está ausente, quando esta de facto está ausente.

Estas duas definições conduzem a outras duas directamente relacionadas, a *fracção de falsos positivos* e a *fracção de falsos negativos,* dadas por :

$$FFP = \frac{\text{n}^o \text{ de decisões falsas positivas}}{\text{n}^o \text{ de casos realmente negativos}}$$

e

$$FFN = \frac{\text{n}^o \text{ de decisões falsas negativas}}{\text{n}^o \text{ de casos realmente positivos}}$$

Note-se que estas fracções representam, respectivamente, as fracções de

casos designados por *realmente negativos* e as fracções de casos designados por *realmente positivos* que são decididos incorrectamente.

Se se assumir que todos os casos podem ser diagnosticados como positivos ou negativos (no que diz respeito a uma determinada doença), então, o número de decisões correctas mais o número de decisões incorrectas deverá ser igual ao número de casos com esse estado actual.

Assim, verifica-se que:

$$FVP + FFN = 1$$

e

$$FVN + FFP = 1.$$

A figura 2.6, pretende explicitar a relação entre o *valor de corte* e a definição dessas fracções, sendo claro que diminuir a *FFP* conduz a um aumento de *FFN*.

Em geral, um teste de diagnóstico tende a ser avaliado por duas destas medidas, *FVP (sensibilidade)* e *FVN (especificidade)*. Metz [58], define *sensibilidade* como sendo a probabilidade de decidir se a doença em questão está presente quando de facto está presente, e *especificidade* como sendo a probabilidade de decidir se a doença em questão está ausente quando, de facto está ausente. Em termos de diagnóstico, poder-se-á definir *sensibilidade* como a capacidade que um teste tem para detectar a doença no indivíduo, e a *especificidade* como a capacidade que o teste tem para excluir os indivíduos isentos de doença. Assim, *valores de corte* elevados, conduzem a um teste pouco sensível e muito específico, por outro lado, *valores de corte* baixos, conduzem a um teste muito sensível e pouco específico.

Num teste de diagnóstico as hipóteses podem ser definidas como:

H_0: O indivíduo é anormal, X_A

H_1: O indivíduo é normal, X_N ,

consequentemente:

$$\alpha = Prob(erro\ de\ tipo\ I) = P(rej\ H_0|H_0) = P(T^-|X_A) = \quad (2.7)$$
$$= 1 - P(T^+|X_A) = 1 - sensibilidade$$

$$\beta = Prob(erro\ de\ tipo\ II) = P(aceitar\ H_0|H_1) = P(T^+|X_N) = \quad (2.8)$$
$$= 1 - P(T^-|X_N) = 1 - especificidade$$

Atendendo a que o *valor de corte* define a região de rejeição, isto é, define a dimensão dos erros de *tipo I* e de *tipo II*, à medida que se varia o *valor de corte* estes erros vão variando, existindo um balanço, à medida que α aumenta, β diminui, e vice-versa.

Na prática torna-se desejável ter um teste que seja ao mesmo tempo altamente sensível e altamente específico, pois um *valor de corte* fixa um par *sensibilidade/especificidade*. Estes pares podem ser representados como valores de coordenadas "y" e "x" dando origem ao gráfico designado por *curva ROC*. Este gráfico permite ter uma noção da capacidade de discriminação de um teste, como será visto na secção 2.6.

A representação ROC em termos de diagnóstico, dá a probabilidade de aceitar H_0, isto é, considerar o indivíduo anormal.

2.6 Curvas ROC

2.6.1 Plano Unitário

Por definição, uma curva ROC é a representação gráfica dos pares *sensibilidade* ou *FVP* (ordenadas) e *1- especificidade* ou *FFP* (abcissas), resultantes da variação do *valor de corte* ao longo de um eixo de decisão, x, a representação gráfica assim resultante é designada por *curva ROC no plano unitário.*

Com efeito, uma curva ROC é uma descrição empírica da capacidade do sistema de diagnóstico poder discriminar entre dois estados num universo, onde cada ponto da curva representa um compromisso diferente entre a *FVP* e a *FFP* que pode ser adquirido pela adopção de um diferente *valor de corte* de anormalidade ou nível crítico de confiança no processo de decisão [58].

Sob o ponto de vista da teoria de testes de hipóteses estatísticas, uma curva ROC é conceptualmente equivalente a uma curva que mostra a relação entre a potência de teste e a probabilidade de cometer um *erro de tipo I* com a variação do "*valor crítico*" (*valor de corte)* do teste estatístico [58].

Consoante os critérios adoptados poder-se-á fazer corresponder um ponto na curva ROC. Assim, pode-se definir, um *critério "estrito"* (por exemplo, apenas se designa o paciente positivo quando a evidência da doença é muito forte) como sendo aquele que conduz a uma pequena fracção de falsos positivos e também a uma relativamente pequena fracção de verdadeiros positivos, isto é, gera um ponto na curva ROC que se situa no canto inferior esquerdo do espaço ROC. Progressivamente critérios menos estritos conduzem a maiores fracções de ambos os tipos, isto é, pontos colocados no canto superior direito da curva no espaço ROC. Esta situação pode ser descrita graficamente pela curva ROC apresentada na figura 2.7.

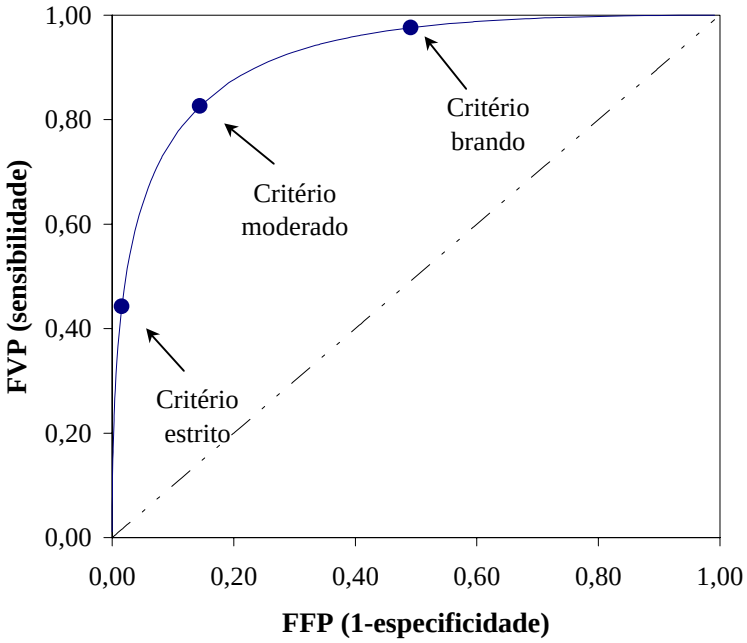


Figura 2.7: Curva ROC, para uma dada capacidade de discriminação, com a variação do critério de decisão.

No que diz respeito ao desempenho de diferentes sistemas de diagnóstico, e considerando a situação em que as curvas ROC associadas a dois sistemas de diagnóstico distintos não se cruzam, o sistema com a curva ROC mais próxima do canto superior esquerdo, fornece um maior poder discriminante. Na figura 2.8, apresentam-se três graus de discriminação possíveis fornecidos pelas curvas ROC.

Quando as curvas ROC se cruzam então podem-se classificar os sistemas para um conjunto de fracções de falsos positivos ou verdadeiros positivos de interesse no sentido do diagnóstico, tendo em conta os custos e benefícios de um diagnóstico alternativo, se necessário [58].

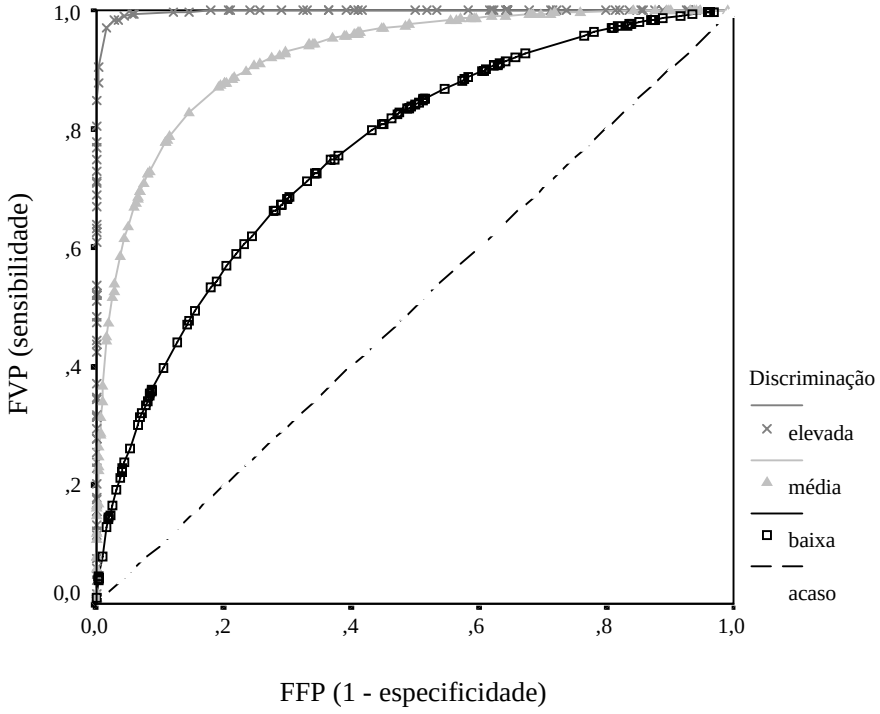


Figura 2.8: Curvas ROC representativas de três graus de capacidade de discriminação.

2.6.2 Plano binormal

Existe uma outra forma de visualizar a curva ROC, através da representação no *plano binormal*, que é um gráfico cujas coordenadas usuais de probabilidade são reescaladas de forma a que os valores dos desvios à Normal sejam linearmente espaçados.

A forma de representar os dados da ROC no plano binormal é através do papel de probabilidades normal [33]. A escala de probabilidades é construída fazendo o cálculo do valor z, para cada valor de P, de acordo com a equação:

$$P(z) = \frac{100}{\sqrt{2\,\pi}} \int_{-\infty}^{z} \exp\left(-\frac{x^2}{2}\right)$$

Assim, o papel de probabilidades normal usa escalas de probabilidade para cada um dos seus eixos (figura 2.9).

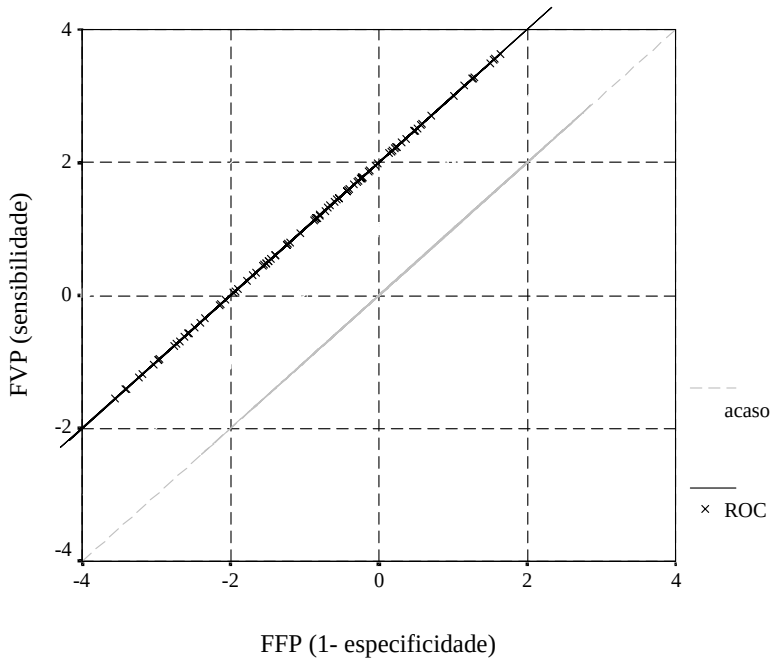


Figura 2.9: Curva ROC no *plano binormal*.

A vantagem deste tipo de gráfico é que a curva ROC para distribuições Normais é uma linha recta e a separação entre as médias das duas distribuições pode ser retirada deste gráfico, como função da diferença entre o valor da ordenada e da abcissa. A ilustração de uma representação deste tipo encontra-se na figura 2.9.

Metz [58] refere que, "de um modo geral uma curva ROC, é especificada assumindo que esta segue uma forma particular com um ou mais parâmetros ajustáveis. (...) A forma funcional *binormal* para a curva ROC é utilizada

muito frequentemente e verifica-se que fornece bons ajustes às curvas ROC empíricas, medidas numa grande variedade de situações".

No entanto Swets (1996) [80], refere que nem todos os dados *no plano binormal* se ajustam a uma linha recta, acrescentando que de facto isto representa uma dificuldade. Este autor menciona também, que "um desvio à linearidade viola a hipótese da Normalidade, e um declive não unitário viola a hipótese da igualdade de variâncias".

Segundo Metz [58], a curva ROC *binormal* pode ser interpretada, em termos de uma variável de decisão x, proveniente de duas densidades Gaussianas, em que

$$f(x|h_0) = \frac{1}{\sqrt{2\pi\sigma_N^2}} \exp\left(-\frac{(x-\mu_N)^2}{2\sigma_N^2}\right) \tag{2.9}$$

designa a função densidade de probabilidade para os casos designados por *normais*, e

$$f(x|h_1) = \frac{1}{\sqrt{2\pi\sigma_A^2}} \exp\left(-\frac{(x-\mu_A)^2}{2\sigma_A^2}\right) \tag{2.10}$$

designa a função densidade de probabilidade para os casos designados por *anormais*. Considera ainda que

$$X_N \backsim N(0,1) \qquad \text{e} \qquad X_A \backsim N\left(\frac{a}{b}, \left(\frac{1}{b}\right)\right)$$

isto é, que a forma funcional *binormal* para a curva ROC pode ser expressa pelo par de equações:

$$FFP(c) = \Phi(-c) \tag{2.11}$$

e

$$FVP(c) = \Phi(a - b\ c) \tag{2.12}$$

onde Φ é a distribuição cumulativa da Normal padrão, e os parâmetros a e b determinam a curva ROC, e c determina um ponto particular da curva [58].

Para demonstrar este resultado considere-se que X_N é a designação para os valores da variável de decisão para os indivíduos considerados *normais*, e X_A é a designação para os valores da variável de decisão para os indivíduos considerados *anormais,* num teste de diagnóstico cuja variável de decisão é x. Hipoteticamente, pode-se descrever a situação através do gráfico da figura 2.10.

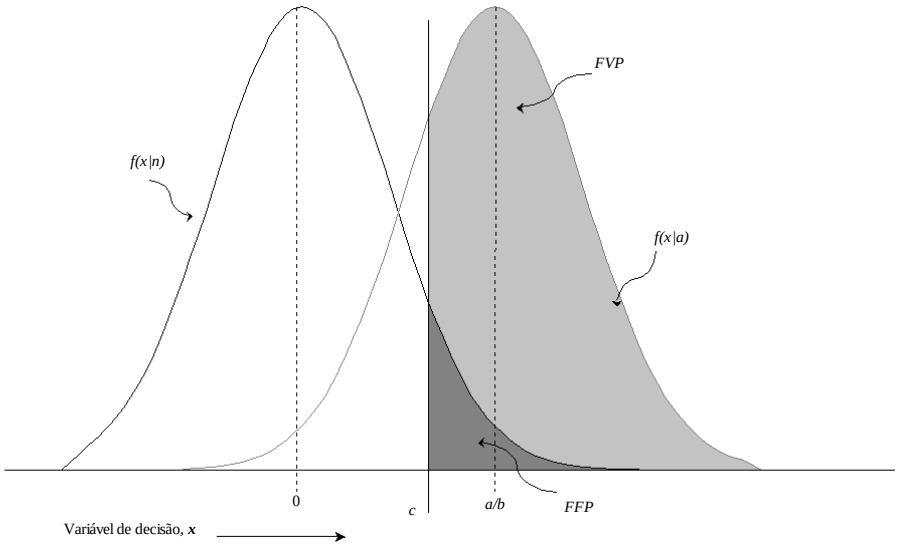


Figura 2.10: Funções de densidade de probabilidade Gaussianas, para os casos designados normais (N) e para os casos designados anormais (A).

$$FFP(c) = \int_c^{+\infty} f(x \mid n)\, dx = \Phi(+\infty) - \Phi\left(z_n\right)_c$$
$$= 1 - \Phi\left(\frac{c-0}{1}\right) = 1 - \Phi\left(c\right) = \Phi(-c) \qquad c.q.d.$$

$$FVP(c) = \int_c^{+\infty} f(x \mid a)\, dx = \Phi(+\infty) - \Phi\left(z_a\right)_c = 1 - \Phi\left(\frac{c-\mu_a}{\sigma_a}\right). \quad (2.13)$$

Por outro lado da equação (2.12), vem:

$$\Phi(a - b\ c) = \Phi\left[-\left(b\ c - a\right)\right] = 1 - \Phi\left(b\ c - a\right) = 1 - \Phi\left(\frac{c - a/b}{1/b}\right). \quad (2.14)$$

Da comparação das equações (2.13) e (2.14), resulta que:

$$\mu_a = \frac{a}{b}$$
$$\sigma_a = \frac{1}{b} \qquad c.q.d.$$

A vantagem da representação *binormal*, como já se referiu, é que a curva ROC é uma linha recta. Num capítulo posterior procurar-se-á demonstrar qual a forma funcional *binormal* para a curva ROC quando são consideradas outras *funções densidade de probabilidade* que não a Normal, e determinar, para o caso das Normais, o significado dos parâmetros a e b.

2.6.3 Índices de precisão das curvas ROC

Existe uma variedade de índices que foram propostos para especificar e/ou resumir as curvas ROC empíricas ([58], [80]). Designa-se por *índice* ou *conjunto de índices*, aquele que especifica uma curva ROC se essa curva puder

ser reconstituída na totalidade a partir de um valor (ou valores) conhecido do índice (ou índices).

As fracções de verdadeiros positivos e falsos positivos, como definidos anteriormente são índices que representam dois tipos de precisão de diagnóstico, e podem ser utilizados para representação da curva ROC.

Também como já referido, a fracção de verdadeiros positivos (FVP) designa o mesmo que *sensibilidade*, isto é, corresponde à proporção de positivos correctamente identificados. Por outro lado, a fracção de verdadeiros negativos (FVN) designa o mesmo que *especificidade*, que corresponde à proporção de negativos correctamente identificados.

Existem outros índices que podem ser utilizados para sumariar uma curva ROC [58], como por exemplo o valor da fracção de verdadeiros positivos num ponto de referência de fracção de falsos positivos, $FVP(FFP_0)$, e o índice área abaixo da curva ROC, A_Z.

2.6.4 Área abaixo da curva ROC

Como mencionado por vários autores ([37], [58], [80]), a área abaixo da curva ROC é um dos índices mais utilizados para sumariar a "*qualidade*" da curva.

De entre os métodos para cálculo de áreas abaixo de uma curva ROC, podem ser considerados os seguintes:

(i) regra do trapézio;

(ii) estimação de máxima verosimilhança, [38], (ver anexo A);

(iii) a partir do declive e termo de intercepção da representação dos dados originais em papel de probabilidades *binormal*, [58] [47];

(iv) aproximação à estatística U de Wilcoxon-Mann-Whitney [38].

Como resultado dos métodos propostos para o cálculo da área abaixo da curva ROC, os erros padrão associados a esta, podem ser obtidos de três maneiras:

(i) como resultado da estimativa de máxima verosimilhança [38];

(ii) a partir da variância da estatística de Wilcoxon (apêndice B);

(iii) como resultado da aproximação à estatística U de Wilcoxon-Mann-Whitney [38].

Capítulo 3

Estado da Arte

Neste capítulo procurou-se fazer uma compilação de algum do material pesquisado sobre o tema da análise *ROC (Receiver Operating Characteristic)*. Serão descritas sumariamente as referências bibliográficas utilizadas no domínio da análise ROC.

3.1 Revisão bibliográfica

O maior contributo para o desenvolvimento da teoria da análise ROC foi dado pela *teoria de detecção de sinal* e também por experiências realizadas no campo da psicologia.

Dos pioneiros no estudo das estimativas de máxima verosimilhança dos parâmetros da *teoria de detecção de sinal*, salienta-se o trabalho desenvolvido por Dorfman e Alf (1969) com o método - *rating-method* - para determinação destas estimativas e respectivos intervalos de confiança [27]. Neste trabalho, os autores, com base no modelo de Thurstone com dois estímulos associados, desenvolvem uma solução para este caso, fornecendo também procedimentos para obtenção da matriz de variâncias-covariâncias e intervalos de confiança.

Para a resolução das equações de verosimilhança, apresentam o método de *scoring* como uma modificação do método de Newton-Raphson. Neste método, as segundas derivadas parciais esperadas substituem as segundas derivadas parciais observadas utilizadas no método tradicional de Newton-Raphson. O *método de scoring,* requer um conjunto de aproximações iniciais ou estimativas preliminares dos parâmetros. Referem que este método, quando comparado com o método do gradiente para resolução de problemas deste tipo, apresenta uma convergência mais rápida. Este resultado iria ser posteriormente utilizado num algoritmo para estimação dos parâmetros na análise ROC.

Green e Swets, em 1966 [33], apresentam um estudo sobre o desenvolvimento da teoria de detecção de sinal e a sua aplicação ao campo da psicofísica. Os autores fazem uma introdução à teoria das probabilidades, teoria estatística de decisão, análise de propagação de ondas e técnicas experimentais. Revêem as experiências básicas que suportam a aplicação da teoria de detecção em psicofísica e descrevem aplicações experimentais desta teoria a uma variedade de problemas substanciais em psicologia.

Irvin Pollack e R. Hsieh (1969) [69] descrevem algumas medidas de precisão utilizadas em psicologia, nomeadamente o índice d'_e. Neste artigo, os autores descrevem dois paradigmas básicos em experiências psicofísicas, o das respostas "*Sim-Não*" e o da "*escolha-forçada*", referenciam as demonstrações feitas por Green e Swets (1966) [33] da relação da área abaixo da curva ROC com a percentagem de respostas correctas no procedimento de "*escolha-forçada*" e no de respostas "*Sim-Não*". Salientam a importância do índice área abaixo da curva ROC e o facto de este ser uma medida não-paramétrica e, por conseguinte, não serem necessários pressupostos sobre as distribuições subjacentes aos dados. Referem que o maior problema do uso

desta medida, é o desconhecimento da sua distribuição amostral. Com base em estudos de simulação, obtêm resultados para a variação da área abaixo da curva ROC para amostras independentes e para amostras correlacionadas pressupondo vários tipos de distribuições e fazendo variar as medidas centrais e de dispersão. Não são conclusivos quanto aos seus resultados, mas apresentam discussões bastante importantes quanto à possível variação da distribuição da área abaixo da curva ROC.

Swets e Pickett (1982) [81] estudam a avaliação de sistemas de diagnóstico a partir de métodos provenientes da teoria de detecção de sinal. Este estudo procura conduzir os diversos resultados que surgem na avaliação do desempenho de um sistema de diagnóstico, ao longo de uma vasta gama de situações nas quais estes sistemas são utilizados. Estas situações incluem a medicina, o controlo de qualidade industrial, a inspecção de materiais e máquinas, a estratégia militar, a pesquisa de informação e a investigação criminal. Referem ainda que "*o diagnóstico em qualquer das suas formas pode ser visto como um problema de detecção de sinal e classificação, e a teoria moderna de detecção de sinal fornece os melhores métodos disponíveis para avaliar sistemas de diagnóstico. Estes métodos fornecem um índice válido e fiável da precisão do diagnóstico e, por conseguinte, satisfazem um primeiro objectivo da avaliação.*"

Apresentam também o algoritmo desenvolvido por Dorfman e Alf (programa RSCORE) para obtenção das estimativas de máxima verosimilhança dos parâmetros da teoria de detecção de sinal para o *método de scoring*. Este programa aplica uma variante do método de Newton-Raphson, designado por *método de scoring*. O programa calcula as estimativas preliminares através do método dos mínimos quadrados como valores iniciais para o *método de scoring*.

Dorfman, em 1973 [26], compara a eficiência do programa RSCORE com outras subrotinas alternativas, como por exemplo a STEPIT, que é uma subrotina que envolve um procedimento do tipo escolha-directa.

O contributo de C. E. Metz foi um dos mais significativos para o desenvolvimento da metodologia da análise através de curvas ROC para avaliação de sistemas de diagnóstico no campo da medicina, nomeadamente em técnicas de imagem médica radiológica. Em 1978 [56] mostra como os diversos conceitos associados à análise ROC estão relacionados. Apresenta, também, definições para os termos *sensibilidade* e *especificidade*, *fracções de verdadeiros positivos* e de *falsos positivos*, *fracções de verdadeiros negativos* e de *falsos negativos*. Questiona os conceitos de *"precisão"* e *"exactidão"* como medidas de diagnóstico. Analisa o significado da curva ROC, descreve algumas curvas ROC experimentais e compara-as. Estabelece uma relação entre a análise ROC e a análise da razão custo/proveito de uma tomada de decisão. São introduzidos os conceitos de "*custo médio de diagnóstico*" e "*proveito líquido médio*", para analisar situações de compromisso para diversos tipos de erros de diagnóstico.

Metz num trabalho conjunto ([60], [58]) apresenta uma nova aproximação considerando um modelo "*binormal bivariado",* para testar diferenças significativas entre duas curvas ROC medidas a partir de dados correlacionados, e mostra como esta aproximação pode ser utilizada para delinear três testes estatísticos distintos, um teste de qui-quadrado bivariado aos parâmetros, teste para as fracções de verdadeiros positivos e teste ao índice área abaixo da curva ROC.

Em 1986, este autor apresenta mais um trabalho sobre a análise ROC na imagem radiológica [57], em que é apresentado um conjunto de referências bibliográficas neste domínio e desenvolvido os conceitos de análise ROC no

diagnóstico de imagem médica.

O índice "*área abaixo da curva ROC*" é uma medida muito utilizada para avaliar o desempenho de sistemas de diagnóstico. Trabalhos como o de Bamber [8] revelam uma relação importante entre este índice e a estatística não paramétrica, U, de Wilcoxon-Mann-Whitney, o que foi um contributo significativo para o conhecimento da estatística associada a este.

A relação verificada por Bamber foi aproveitada por Hanley e McNeil (1982) [37] que, baseados no significado de área abaixo da curva ROC, e utilizando as ligações entre vários conceitos estatísticos, desenvolveram técnicas analíticas para explicitar as propriedades estatísticas da curva. Apresentam uma rotina de cálculo para a estatística de Wilcoxon-Mann-Whitney, assim como para os respectivos erros padrão (SE). Os autores salientam ainda a importância da não existência de pressupostos distribucionais para estes cálculos. Neste artigo é apresentada ainda uma forma de determinação da dimensão óptima da amostra (considerando a igualdade entre as dimensões das amostras dos casos normais e anormais). Estes autores abordam também, um método para detectar diferenças significativas entre áreas abaixo de duas curvas ROC a partir da dimensão da amostra, n, e para vários níveis de confiança, $80\%, 90\%$ e 95%.

Em 1983, Hanley e McNeil [38] estendem a análise desenvolvida em [37] a uma classe mais vasta de situações, quando duas ou mais curvas ROC são geradas usando o mesmo conjunto de dados. Nesta situação torna-se necessário a introdução de uma medida de associação entre os dois conjuntos de observações sobre os dados em questão. Neste artigo é apresentada uma tabela que a partir da média dos coeficientes de correlação r_A (coeficiente de correlação não paramétrico para os casos designados anormais) e r_N (coeficiente de correlação não paramétrico para os casos designados normais), e da

média das áreas A_1 e A_2, determina o coeficiente de correlação r entre as áreas das duas curvas ROC. É ainda estabelecida uma relação para o número de indivíduos necessários em casos de experiências com emparelhamento e para amostras independentes.

Ainda em 1983, McNeil e outros [55] apresentam um estudo para duas técnicas de imagem, utilizando a metodologia descrita por Swets e Pickett [81], com uma modificação para dados emparelhados, comparando-a com a metodologia desenvolvida por Hanley e McNeil [38].

Para comparação de duas ou mais curvas ROC para dados correlacionados, DeLong e outros (1988) [22] apresentam uma aproximação não paramétrica, baseando-se na determinação de uma matriz de variâncias-covariâncias para um vector genérico de estatísticas U. Estes autores referem que a tabela de valores de r determinada por Hanley e McNeil [38] apresenta limitações, nomeadamente o facto de ser aplicada somente em situações em que a diferença média das áreas abaixo das curvas ROC a comparar é superior a 0.7. Neste artigo os autores apresentam uma metodologia alternativa, utilizando uma aproximação não paramétrica que explora as propriedades da estatística U de Mann-Whitney. A técnica utilizada para fornecer estimativas consistentes dos elementos da matriz variâncias-covariâncias do vector U é a desenvolvida no método das componentes estruturais de Sen (1960).

Um artigo com aplicação da metodologia de DeLong foi desenvolvido por Rockette (1990) [75], para a comparação de dois conjuntos de dados para sistemas de imagem.

Autores como Hanley, McNeil e DeLong, basearam os seus estudos no índice área abaixo da curva ROC e a sua aproximação à estatística U de Wilcoxon-Mann-Whitney, para comparação de sistemas de diagnóstico. Porém, outros autores desenvolveram alguns estudos com base na teoria de

detecção de sinal, utilizando outros índices de discriminação, como d', d'_e e Δm [69], [28] e [80].

Hanley (1988) [36] apresenta algumas razões da imensa aplicabilidade da curva ROC, baseada nos pressupostos desenvolvidos na abordagem paramétrica da teoria de detecção de sinal, designada por *curva ROC no plano binormal*.

Neste artigo, algumas justificações formuladas por vários autores para o uso da forma *binormal* são citadas, nomeadamente:

- A distribuição Gaussiana é a natural - "*...many of random variables describing natural phenomena may be considered to be the sum of large, relatively constant number of other independent, random variables;...since we often believe that sensory events are composed of multitude of similar, smaller events, the Central Limit Theorem might be invoked to justify the Gaussian assumption" [33] [pág.54-58];*

- Outras distribuições podem ser aproximadas pela Gaussiana - "*...the binomial, Poisson, hypergeometric, and chi-squared distributions can, under certain conditions, be closely approximated by the Normal distribution" [33] [pág. 58];*

- O eixo de decisão pode ser transformado para produzir distribuições Gaussianas- *"...any monotonic transformation of decision-variable axis yields generally different underlying distributions but the same ROC curve" [28];*

- Outras formas ROC parecem "aproximar-se de uma recta" no

papel *binormal* - *"...the plot of Power-Law ROCs in binormal coordinate shows that they are nearly straight lines" [28] ;*

- Resultados empíricos mostram que a forma *binormal* se ajusta a uma recta - *"...it is a highly robust, empirical result, wich is now substantiated in dozen of diverse applications, that empirical ROC is very similar in form to a theoretical ROC derived from normal probability distributions. In practice, in other words, the ROC curve is adequately described by a straight line when plotted on binormal graph" [81] [pág. 5 e 30];*

- Tratabilidade matemática e conveniência - *"...it has the convenient property that all possible binormal ROC curves are transformed into straight lines if plotted on normal deviate axes" [59] citando [33];*

"...it is relatively easy to fit by eye and is easily fitted by statistical techniques that give estimates of the slope and intercept of binormal ROC" [81] [pág. 31].

Hanley [36] acrescenta ainda a esta lista de "popularidade da forma *binormal*", o facto dos programas disponíveis para traçar a curva ROC, utilizarem esta forma por facilidade em termos de cálculo.

O modelo *binormal* foi, sem dúvida, o mais utilizado para descrever sistemas de diagnóstico. Autores como Swets [78], [79] e Ratcliff [71], aplicam esta metodologia em campos associados à psicologia.

Iverson e outros, em 1992 [47], através da generalização das propriedades do modelo padrão da Normal na teoria de detecção de sinal, apresentam definições importantes no desenvolvimento da teoria ROC, como por exemplo

o *"teorema da área"*. Estes autores demonstram também, neste artigo, para uma curva ROC no plano *binormal*, a relação entre a distância da recta ajustada à origem e o valor da área abaixo da curva ROC.

Muitos autores, como Philbrick (1980) [68], Diamond (1986) [23], Hlatky (1987) [41] e Tavel (1987) [83], optaram pelo uso da *sensibilidade* e da *especificidade* como medidas de precisão de diagnóstico apesar das suas limitações.

Diamond (1986) [23] apresenta uma aplicação destes conceitos como medida de estudo de diagnóstico em doenças coronárias. Concluí, no entanto, que a *sensibilidade* e a *especificidade* não deverão ser consideradas suficientes num grupo de referência pequeno.

Philbrick (1980) [68] e Hlatky (1987) [41] apresentam trabalhos com aplicação destes conceitos ao estudo de doenças coronárias, apresentando algumas vantagens do seu uso nos testes de diagnóstico.

No estudo de Hlatky [41], os autores concluíram que a *sensibilidade* e a *especificidade* variam com as características clínicas e, por conseguinte, deveriam ser tidos em conta os factores clínicos no desempenho da análise do teste de diagnóstico.

Também em 1987, Tavel [83] revela a importância do uso da *especificidade* e da *sensibilidade* como medidas de diagnóstico e discute possíveis falhas na sua utilização. Apresenta como solução para o problema da dependência da *sensibilidade* e da *especificidade* das distribuições de respostas negativas e positivas ao teste, a verificação do estudo prospectivamente para todos os pacientes.

Begg (1991) [9] efectua um resumo sobre as metodologias de diagnóstico utilizadas em medicina. Reforça a ideia de que é necessário fazer um correcto planeamento da experiência, no que diz respeito à recolha de dados, escolha de um *"gold standard" (grupo referência)* e metodologia utilizada na análise

estatística. Refere-se à análise ROC como sendo um dos últimos ressurgimentos na análise de diagnóstico e uma boa medida de precisão de um teste de diagnóstico.

John A. Swets, em 1996 [80] apresenta uma colecção de artigos sobre a teoria de detecção do sinal e a análise ROC em psicologia e diagnóstico. Este livro é composto por três grandes blocos; no primeiro bloco é apresentado um conjunto de artigos sobre conceitos teóricos associados à análise ROC e teoria do sinal; no segundo bloco são apresentados dois artigos sobre precisão e eficiência de diagnóstico; por fim, no terceiro bloco, o autor apresenta um conjunto de artigos referentes às aplicações desta teoria nos mais diversos campos.

Em 1996 Halpern e outros [35], num estudo sobre comparação de sistemas de diagnóstico, desenvolveram um novo método para comparação das curvas ROC baseado no que designaram por *pontos óptimos de operação (OOP - Optimal Operating Points)*. Referem que a eficiência do diagnóstico poderá ser avaliada apenas pela comparação de pontos óptimos de operação, para um valor fixo de fracções de verdadeiros positivos ou de fracções de falsos positivos, em alternativa ao índice área abaixo da curva ROC.

Um conjunto de autores como, Swets (1979) [82], Gatsonis (1990) [31], Henkelman (1990) [39], Rifkin [74], Colliver (1992) [18], Mossman (1994) [61], Chen (1994) [17], Parker (1995) [67], Burdette [11], Eskicioglu (1996) [29], Jiang [48], McMillan (1996) [54] e Holmes[42], utilizam a metodologia da análise ROC, quer para determinação de valores óptimos de *valores de corte*, quer para comparações de diferentes testes, nos mais diferentes campos, como por exemplo psicologia, medicina e técnicas de imagem.

Swets (1979) [82] utilizou um protocolo para a avaliação rigorosa de um sistema de diagnóstico em medicina, num estudo comparativo de duas

técnicas radiológicas para detecção, localização e diagnóstico de lesões cerebrais: tumografia computorizada (TC) e scanning radionuclear (RN). Para tal, utilizou as leituras de seis técnicos em TC e outros seis em RN. Traçou as ROC em papel de probabilidades *binormal* para cada um dos seis leitores em separado.

Gatsonis (1990) [31] discute a avaliação de tecnologias clínicas em radiologia, concentrando-se essencialmente em estudos prospectivos, isto é, estudos nos quais os pacientes são recrutados e testados com ferramentas clínicas, em vez de serem seleccionados na base de um estudo retrospectivo. Neste artigo os autores tecem algumas considerações a ter em conta no estudo comparativo de técnicas de imagem radiológica, nomeadamente no que diz respeito ao grupo de investigação, técnicas a analisar, instituições participantes, desenvolvimento de um protocolo para o estudo, implementação do controle de qualidade, metodologia estatística utilizada e pressupostos de utilização.

Para Henkelman (1990) [39], a avaliação de testes de diagnóstico médico requer, tradicionalmente, precisão, avaliação independente do estado da doença do paciente contra o qual o teste pode ser comparado. Salientam que a análise ROC é uma aproximação para avaliação de testes de diagnóstico, que possui técnicas analíticas que permitem traçar um gráfico de FVP (fracções de verdadeiros positivos) versus a FFP (fracção de falsos positivos), a curva ROC, cujos pontos de operação correspondem a pares de sensibilidade/especificidade. A análise ROC é independente de certos factores como a prevalência da doença na população e da escolha do critério de decisão pelo observador. No entanto, a análise ROC continua a requerer uma medida do verdadeiro estado da doença em cada paciente. Descrevem um método de análise que utiliza os dados ROC para comparar a precisão de testes de diagnóstico que apresentam problemas provenientes da necessidade de um "*gold standard*".

O método desenvolvido é aplicável a situações em que as comparações são realizadas para dois ou mais testes, e cada um deles apresenta um elevado nível de precisão (área abaixo da curva ROC acima de 0.9). O método pode também ser aplicado a situações em que existe "*gold standard*", para avaliar a consistência dos dados. O método produziu resultados equivalentes à análise ROC convencional na comparação do TAC (Tomografia Axial Computorizada), da ressonância magnética e de RS (Cintigrafia Radionuclear) obtida para metastases vivas.

Neste artigo é apresentado um comentário de Metz, que salienta que o objectivo do estudo de Henkelman foi comparar dois ou mais testes radiológicos quando não existe "*gold standard*" nem "*consensus diagnosis*" dos testes em estudo. Os seus métodos empregam o uso de mistura de distribuições, isto é, os dados são gerados a partir (neste caso) de duas populações, doentes e não doentes, mas onde o verdadeiro estado de doença é desconhecido. No seu exemplo, são comparados três testes radiológicos, sendo cada um testado nos cinco pontos usuais de uma escala de classificação. As três variáveis lactentes contínuas correspondentes a estas classificações foram assumidas como sendo geradas a partir da mistura de duas distribuições normais trivariadas, tal como na análise ROC convencional. Os parâmetros destas distribuições e a proporção de mistura (ou prevalência da doença) foram estimadas utilizando um algoritmo iterativo que deu as estimativas de máxima verosimilhança. Os autores comparam os resultados desta análise com a análise convencional onde o verdadeiro diagnóstico é conhecido, e provaram uma certa consistência. No entanto, o uso desta metodologia tem sérias e potenciais limitações. Uma importante limitação é que os parâmetros das distribuições componentes podem não ser na realidade identificáveis, especialmente se as distribuições estiverem perto uma da outra. Um outro problema potencial

é causado pelo facto de num conjunto radiológico não se disporem de dados contínuos, apenas resultados numa escala com cinco classificações.

Rifkin [74], através da utilização do programa CORROC2, específico para análise ROC para duas amostras correlacionadas, desenvolvido por Metz, compara duas técnicas de imagem, a ressonância magnética (MRI) e a ultrasonografia (US), na detecção do cancro da próstata em diversos estádios da doença. Este estudo permitiu-lhes concluir, que nenhuma das técnicas avaliadas, se mostrou eficiente na detecção precoce deste tipo de doença.

Também Goddard e outros (1990) [32] avaliam o desempenho de alguns "*kits*" de diagnóstico para determinar os níveis de "*serum prostatic acid phosphatase*" em pacientes com diferentes estádios de cancro na próstata, sendo cada paciente estudado com vários "*kits*". Comparam os resultados obtidos através da metodologia da curva ROC, assumindo que os dados seguem uma distribuição Normal, que a transformação logarítmica dos dados segue uma Normal, e nenhum tipo de distribuição associada aos dados. Verificaram diferenças importantes entre os resultados das diferentes aproximações. Afirmam que para este conjunto de dados, a aproximação à Normal deverá ser utilizada com extrema precaução. A transformação logarítmica dá resultados que são comparados favoravelmente com os da não paramétrica, mas uma aplicação irreflectida do método deveria ser evitada.

No estudo de Colliver e outros (1992) [18] utilizaram-se critérios quantitativos, assim como considerações práticas para determinar os valores óptimos para o comprimento do *teste de screen* (isto é, o número de casos) e a localização do *valor de corte*. Utilizaram-se as curvas ROC para vários *testes de screen*, variando o comprimento, onde os pontos em cada curva correspondem a diferentes valores de corte no *teste de screen*. Os resultados demonstraram que pode ser obtida uma boa precisão com um *teste de screen* que contenha

apenas uma terça parte do comprimento total, e o *valor de corte* para este teste deveria ser ligeiramente acima da média dos níveis que maximizam a *sensibilidade* e a *especificidade*. Referem também que um indicador quantitativo comum da precisão descrita por uma curva ROC, a área abaixo da curva, pode ser interpretado como sendo a probabilidade de uma classificação correcta de pares examinados. Calcularam a área abaixo da curva ROC através da regra do trapézio.

Mossman (1994) [61] recomenda o uso da análise ROC para avaliar o efeito de detecção ou previsão da violência. O autor afirma que os métodos ROC descrevem a exactidão dos índices que não são afectados por classificações de base ou pelos enviesamentos clínicos a favor ou contra a previsão dos erros de tipo I ou de tipo II. Os métodos ROC ocupam uma posição central ou unificada nos processos de determinação e uso de ferramentas de diagnóstico na medicina clínica e foram utilizados para avaliar a previsão da delinquência juvenil. Numa primeira análise, o autor explica como a análise ROC pode auxiliar os investigadores a descrever e avaliar a previsão da violência. Considera a construção da curva ROC assente no pressuposto de que a distribuição é normal bivariada. Os autores utilizaram um software denominado ROCFIT que dá os índices da análise ROC.

Chen e outros (1994) [17] procuram utilizar a análise ROC para estudar a exactidão do CBCL (Child Behavior Checklist) para testes de deficiência de atenção de desordem hiperactiva (ADHD) [crianças com ADHD têm valores elevados de CBCL]. Os autores comparam esta escala para quatro grupos, com outras escalas diferentes, e concluem que a CBCL é a melhor para qualquer destes grupos. Referenciam a análise ROC de qualidade (QROC) como sendo uma possível transformação para a análise ROC, utilizando como índices os valores de *kappa* e d_Q (distância de cada sintoma do ponto ideal).

Concluem que à medida que a eficiência do diagnóstico aumenta, o valor de d_Q diminui.

Parker (1995) e outros [67], em estudos de cancro da mama, utilizaram a técnica da análise ROC, através do índice área abaixo da curva ROC, para avaliação de um sistema de classificação. Testaram um conjunto de 42 casos onde se obteve um valor para a área abaixo da curva ROC de 0.91 usando uma combinação de seis tipos de *clusters de calcificação*.

Em Burdette (1996) [11], através da utilização do programa CORROC2, é feita a comparação de duas técnicas de diagnóstico da doença de Alzheimer.

Eskicioglu (1996) [29] desenvolve uma técnica de melhoramento da imagem sem ter de recorrer às ferramentas padrão para medir a qualidade de imagens reconstruídas, como é o caso do método da média quadrática do erro normalizada (NMSE). O autor considera também, a análise ROC como uma possível ferramenta para a medida de qualidade da imagem, mas que no entanto, envolve muitos custos e demora de tempo. Refere ainda que devido a estes factores, a análise ROC se torna demasiado específica para cobrir uma gama de modalidades no campo da imagem médica e suas aplicações.

Jiang e outros (1996) [48], utilizam a metodologia ROC, usando o programa LABROC4, para classificação e comparação de uma técnica computorizada de detecção de microcalcificações benignas ou malignas (origem do cancro da mama), com a técnica usual dos radiologistas.

McMillan e outros (1996) [54] pretendem determinar o desempenho clínico de três tipos de fórmulas de fluídos cerebroespinais, (CSF) IgG, utilizando os dados obtidos a partir de dois métodos quantitativos. Os métodos usados foram os da análise ROC e o dos índices de decisão. Foram traçados gráficos para comparar a RN (rate nephelometric) e a RIEP (rocket immunoelectrophoretic). Estas fórmulas foram utilizadas para determinar o teste

clinicamente mais preciso para o diagnóstico da esclerose múltipla, tendo em conta a sua precisão e o custo efectivo da análise. Utilizando o método RN com uma determinada concordância, para um valor de corte de 90% de *especificidade*, o índice IgG dá melhor desempenho clínico. Os autores concluíram que a curva ROC e a análise dos gráficos dos índices de decisão, fornecem ferramentas valiosas na determinação e comparação do desempenho clínico de testes laboratoriais novos e dos já existentes.

Ainda no campo da medicina, para o estudo dos índices de avaliação de risco neonatal inicial, The International Neonatal Network (1993) [64] determinou que o CRIB é um índice robusto para determinar o risco neonatal inicial e mais preciso do que o peso à nascença.

Em 1994, Rautonen e outros [72] utilizaram também a metodologia ROC para avaliar o risco de morte para recém-nascidos prematuros através de três índices: CRIB, SNAP e SNAP-PE. Através das curvas ROC procuraram determinar qual o melhor índice para a previsão do risco de morte.

Courcy-Wheeler e outros (1995) [21], num estudo prospectivo para recém-nascidos com muito baixo peso à nascença (<1500 g) e prematuros (< 32 semanas), procuraram determinar a capacidade do CRIB face à idade gestacional, para prever as taxas de mortalidade e morbilidade e ainda os tempos de permanência nas Unidades de Cuidados Intensivos Neonatais, utilizando para o efeito estudos através da análise ROC.

Após a apresentação da pesquisa bibliográfica efectuada no enquadramento do tema desta dissertação, pode-se verificar, por um lado, a existência de muitos contributos neste domínio, por outro, a grande diversidade de aplicações da análise ROC.

Este capítulo pretende também mostrar como alguns trabalhos desenvolvidos em domínios como a teoria de detecção de sinal, teoria estatística e

análise de diagnóstico, serviram de ponto de partida para o surgir da análise ROC. Assim, sob uma perspectiva cronológica, e atendendo à evolução da teoria da análise ROC, poder-se-à agrupar a informação recolhida em dois blocos básicos:

- o da teoria de detecção de sinal que conduz a uma abordagem designada por paramétrica, pois os pressupostos em que esta assenta são os da Normalidade das distribuições, modelo apresentado por Thurstone referenciado em [80];

- o desenvolvido em torno da aproximação do índice área abaixo da curva ROC à estatística não paramétrica de Wilcoxon-Mann-Whitney, iniciado em estudos como o de Bamber [8], que conduz à abordagem não paramétrica.

Verifica-se ainda que a maioria das aplicações da metodologia ROC utiliza a abordagem paramétrica, o que pode ser devido, entre outros factores para além dos enunciados por Hanley em [36], à existência de programas que utilizam apenas este tipo de abordagem.

Capítulo 4

Principais contributos para o desenvolvimento da análise ROC

Este capítulo procurará dar a conhecer quais foram os principais contributos para o desenvolvimento da análise *ROC (Receiver Operating Characteristic)*. Assim, apresentar-se-á o resumo de alguns trabalhos que contribuíram de uma forma significativa para o desenvolvimento da teoria da análise *ROC*, realizados por diferentes autores, assim como algumas relações importantes.

4.1 Relação entre a área abaixo da curva ROC e a área do *Gráfico de Ordenação Dominada*

Em 1975, os gráficos *ROC* foram interpretados por Bamber [8], como uma variante da designada curva de *Ordenação Dominada - Ordinal Dominance*

Curve (OD).

O gráfico de *Ordenação Dominada* (*Ordinal Dominance Graph)* para as variáveis X e Y, ou (X, Y) OD foi definido por Darlington (1975) [8]. Dadas duas variáveis aleatórias X e Y, e c uma constante arbitrária, considere-se o gráfico cujo eixo coordenado horizontal é dado pela probabilidade $P(X \leq c)$, e o eixo coordenado vertical é dado pela probabilidade $P(Y \leq c)$. Designe-se por $T(c)$, o ponto traçado neste gráfico, para todos os valores possíveis de c, desde $-\infty$ a $+\infty$. O gráfico assim resultante é designado por gráfico de *Ordenação Dominada* de uma população, como exemplificado na figura 4.1.

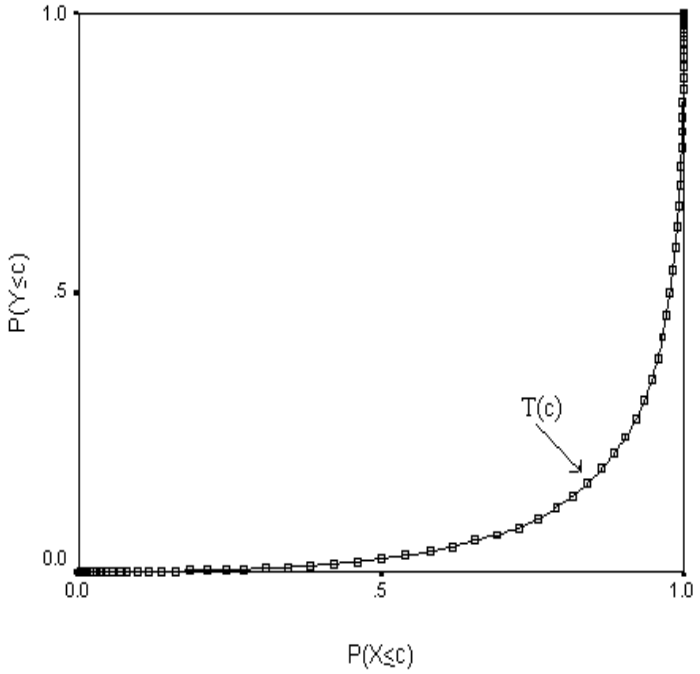


Figura 4.1: Gráfico de Ordenação Dominada (OD) de uma população.

Uma propriedade deste tipo de gráficos é que são invariantes no sentido de preservação de transformações, isto é, dada uma função m estritamente

crescente, definida para todos os valores das variáveis aleatórias X e Y, então o gráfico OD para X e Y é idêntico ao gráfico OD para $m(X)$ e $m(Y)$.

Bamber [8] desenvolveu o seu trabalho considerando as variáveis aleatórias X e Y contínuas, ou então discretas finitas, apresentando as definições de gráfico OD para os dois casos.

Citando este autor, segundo Birnbaum e Klose (1957), uma variável aleatória X diz-se estocasticamente menor ou igual do que outra variável aleatória Y, para qualquer constante c, se:

$$P(X \leq c) \geq P(Y \leq c).$$

Bamber afirma ainda que duas variáveis aleatórias X e Y dizem-se estocasticamente comparáveis quer no caso de X ser estocasticamente menor ou igual do que Y, quer no caso contrário, Y ser estocasticamente menor ou igual do que X. Se se considerar a *diagonal positiva* ($P(X \leq c) = P(Y \leq c)$), como a linha que une os pontos $(0,0)$ e $(1,1)$, no plano onde se encontra traçado o gráfico OD, então duas variáveis aleatórias X e Y dizem-se estocasticamente comparáveis se e só se a curva OD se encontra, por completo, abaixo da *diagonal positiva*.

Bamber designa área acima do gráfico OD para X e Y por $A(X,Y)$. Se X e Y forem contínuas, e se f_Y designar a função densidade de probabilidade de Y, então:

$$\begin{aligned} A(X,Y) &= \int_0^1 P(X \leq c)\ dP(Y \leq c) \\ &= \int_{-\infty}^{+\infty} P(X \leq c)\ f_Y(c)\ dc \\ &= P(X \leq Y). \end{aligned}$$

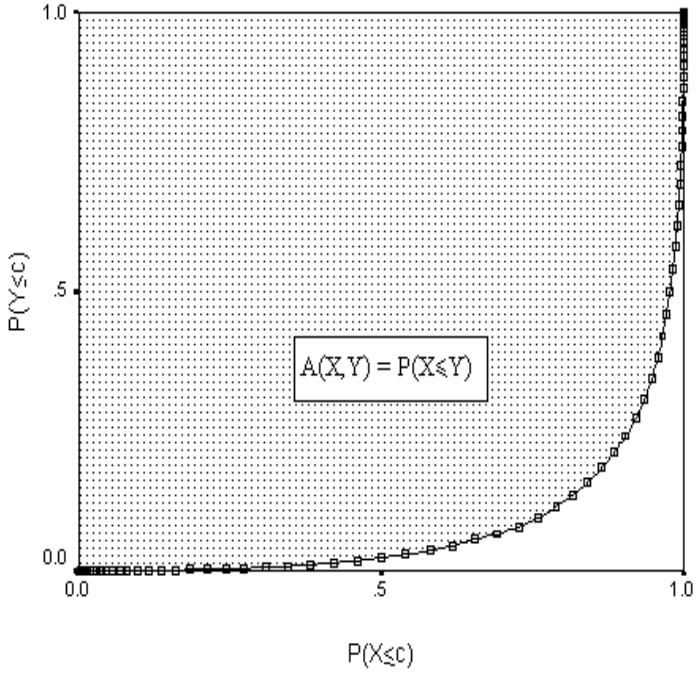


Figura 4.2: Área acima do gráfico OD, para X e Y contínuas.

que corresponde à área sombreada na figura 4.2.

Para X e Y discretas finitas considera-se $c_1, ..., c_k$ o conjunto ordenados de valores que X e Y podem tomar com probabilidade não nula. Seja ainda, c_0 um valor arbitrário menor que c_1. A área acima do gráfico OD pode ser calculada dividindo esta em trapézios, e calculando a área de cada trapézio. Considere-se, assim, A_i a área do trapézio considerado na figura 4.3.

Analiticamente, poder-se-à calcular a área de cada trapézio através de:

$$\begin{aligned} A_i &= P(Y = c_i) \left[\frac{1}{2} P(X \leq c_i) + \frac{1}{2} P(X \leq c_{i-1})\right] \\ &= P(Y = c_i) \left[P(X \leq c_{i-1}) + \frac{1}{2} P(X = c_i)\right]. \end{aligned}$$

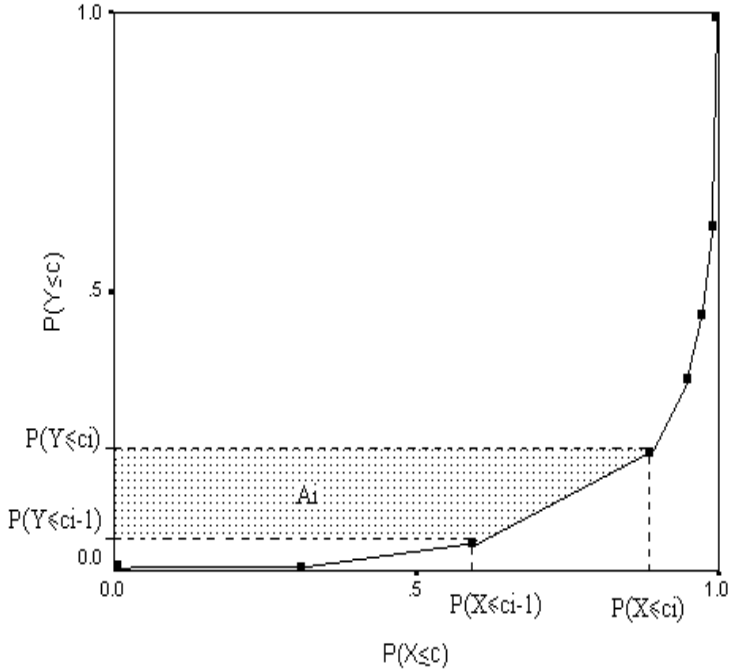


Figura 4.3: Área acima do gráfico OD, para X e Y discretas finitas.

Como $A(X,Y)$ é igual à soma dos diversos A_i, vem:

$$\begin{aligned} A(X,Y) &= \sum_{i=1}^{k} P(Y=c_i)\ \left[P(X \leq c_{i-1}) + \frac{1}{2}\ P(X=c_i)\right] \\ &= \sum_{i=1}^{k} P(Y=c_i)\ P(X \leq c_{i-1}) + \frac{1}{2}\sum_{i=1}^{k} P(Y=c_i)\ P(X=c_i) \\ &= P(X<Y) + \frac{1}{2}\ P(X=Y) \end{aligned} \tag{4.1}$$

Dado que para X e Y contínuas, $P(X=Y)$ é nula, então a equação (4.1) é válida para ambas as situações, variáveis aleatórias discretas finitas e variáveis aleatórias contínuas.

A partir da equação (4.1) nota-se que a medida de área $A(X,Y)$ avalia

a distância à qual a distribuição de X se encontra da distribuição de Y, em termos proporcionais. Então, $A(X,Y)$ pode tomar qualquer valor desde um mínimo igual a zero até um máximo igual a um. O valor máximo é obtido se e só se a distribuição de X compreender valores, por completo, abaixo dos da distribuição de Y sem existência de sobreposição das duas distribuições. De forma análoga $A(X,Y) = 0$, se e só se a distribuição de X compreender valores por completo, acima dos da distribuição de Y sem existência de sobreposição das duas distribuições. Por outro lado, se as duas distribuições forem identicamente distribuídas, isto é, se apresentarem uma sobreposição completa, então $A(X,Y) = \frac{1}{2}$. De salientar ainda que $A(X,Y)$ e $A(Y,X)$ são complementares, pelo facto de que a sua soma é sempre um.

Bamber [8] afirma que as propriedades de $A(X,Y)$ tornaram-na uma medida útil, da dimensão ou importância, da diferença entre duas populações.

Da mesma forma que se definiu um gráfico OD de população, pode-se também definir um gráfico OD para uma amostra, sem perda de generalidade dos conceitos introduzidos. Assim, considere-se uma amostra aleatória com N_X observações da variável aleatória X, e uma amostra aleatória com N_Y observações da variável aleatória Y. Seja $p(X \leq c)$ a *proporção* de observações de N_X de X que são menores ou iguais que uma constante c, e $p(Y \leq c)$ a *proporção* de observações de N_Y de Y que são menores ou iguais que uma constante c. Para cada c, seja $t(c)$ o ponto de coordenada horizontal $p(X \leq c)$ e coordenada vertical $p(Y \leq c)$, então um gráfico OD amostral para X e Y é formado por pontos $t(c)$ para todo c de $-\infty$ a $+\infty$. Para qualquer c, cada coordenada do ponto $t(c)$ é um estimador não enviesado para as correspondentes coordenadas de $T(c)$. Neste sentido, o gráfico OD amostral para X e Y, pode ser considerado um estimador não enviesado do gráfico OD da população de X e Y.

Dada uma amostra aleatória com N_X observações da variável aleatória X, e uma amostra aleatória com N_Y observações da variável aleatória Y, então existe um total de $N_X.N_Y$ combinações possíveis de X com Y. Designe-se por $p(X < Y)$, $p(X = Y)$ e $p(X \neq Y)$ a proporção dos pares N_X N_Y para os quais $X < Y$, $X = Y$ e $X \neq Y$, respectivamente. Então os estimadores não enviesados de $P(X < Y)$, $P(X = Y)$ e $P(X \neq Y)$ são $p(X < Y)$, $p(X = Y)$ e $p(X \neq Y)$, respectivamente. Se designar por $a(X, Y)$ a área acima do gráfico OD amostral, então pela regra do trapézio, semelhante à equação (4.1), virá:

$$\begin{aligned} a(X,Y) &= \sum_{i=1}^{k} p(Y = c_i) \ \left[p(X \leq c_{i-1}) + \frac{1}{2}\, p(X = c_i) \right] \\ &= \sum_{i=1}^{k} p(Y = c_i)\ p(X \leq c_{i-1}) + \frac{1}{2} \sum_{i=1}^{k} p(Y = c_i)\ p(X = c_i) \\ &= p(X < Y) + \frac{1}{2}\, p(X = Y). \end{aligned} \tag{4.2}$$

Para verificar que $a(X, Y)$ é um estimador não enviesado para $A(X, Y)$, basta tomar o valor esperado em ambos os membros da equação (4.2):

$$\begin{aligned} E\left[a(X,Y)\right] &= E\left[p(X < Y) + \frac{1}{2}\, p(X = Y) \right] \\ &= E\left[p(X < Y)\right] + \frac{1}{2}\, E\left[p(X = Y)\right] \\ &= P(X < Y) + \frac{1}{2}\, P(X = Y) \end{aligned}$$

donde resulta,

$$E\left[a(X,Y)\right] = A(X,Y).$$

Assim, a área acima do gráfico OD amostral é um estimador não enviesado da área acima do gráfico OD da população.

Bamber [8] refere-se também à existência de uma relação entre o estimador $a(X,Y)$ e a estatística U de Mann-Whitney. Dado que a estatística U é definida como sendo o número total de pares (X,Y) para os quais $X < Y$, então se X e Y forem contínuas,

$$a(X,Y) = \frac{U}{N_X \; N_Y}.$$

Este é um resultado importante, na medida que permite maior simplicidade no cálculo desta grandeza.

A curva ROC pode então ser visualizada como uma variante do gráfico OD, por rotação deste, isto é, a curva ROC pode ser vista como um gráfico cujo eixo coordenado vertical corresponde a $P(Y \geq c)$ e o eixo coordenado horizontal corresponde a $P(X \geq c)$, como exemplificado na figura 4.4.

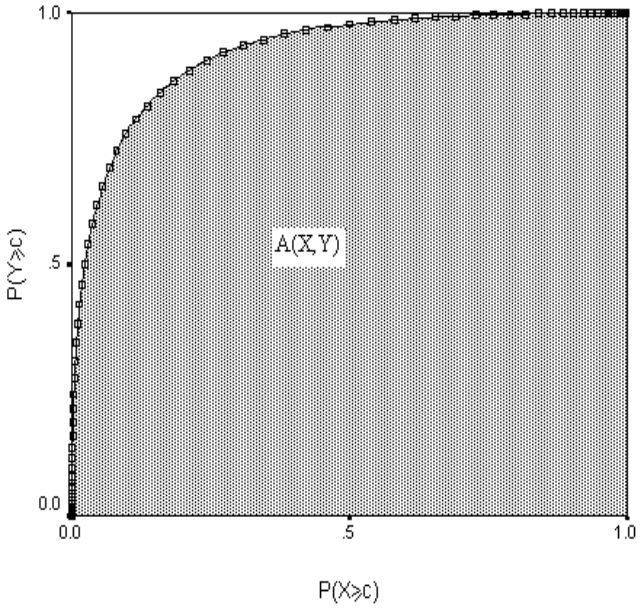


Figura 4.4: Um exemplo de curva ROC.

A partir deste gráfico, também é possível observar que a área acima do gráfico *OD* corresponde à área abaixo da curva *ROC*.

4.2 Procedimento de resposta *"sim-não"*

Green e Swets [33] utilizaram a metodologia da teoria de detecção de sinal em problemas de decisão no campo da psicologia. Pode-se dizer que a componente principal da teoria de detecção é a aplicação da teoria de decisão a situações nas quais aos *"sinais"* pode ou não ser adicionada uma perturbação aleatória, o *"ruído"*.

No designado procedimento de resposta *"sim-não"*, o observador do acontecimento responde *"sim"* se pensar que o sinal está presente nessa experiência, e responde *"não"* no caso contrário. Neste tipo de procedimento assume-se que o observador decide a sua escolha baseado num critério e, para cada experiência, se a sua leitura (impressão) excede o critério pré-adoptado, ele responde *"sim"*. Poder-se-à representar uma sequência típica de acontecimentos associados a este procedimento de detecção binária como exemplificado na figura 4.5.

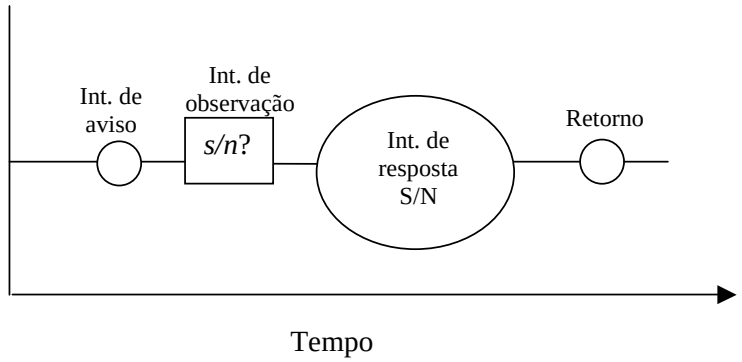


Figura 4.5: Acontecimentos numa experiência de procedimento "*sim-não*".

Neste tipo de procedimento designa-se por s e n as duas alternativas de estímulo, que correspondem respectivamente, presença de *sinal* e *ruído*. As respostas possíveis do observador irão ser designadas por S (*"sim" sinal*) e N (*"não sinal", ruído*). Não existe outro tipo de resposta, nomeadamente o observador não pode responder *"não sei"*.

Como existem apenas dois graus de liberdade neste procedimento, isto é, basta o conhecimento de duas probabilidades para se conhecerem as outras duas, dado que a representação do acontecimento *"estímulo-resposta"*, se pode resumir a uma tabela 2×2, como referido no capítulo anterior.

Toda a informação contida nesta tabela se pode resumir a um ponto num gráfico a duas dimensões, cujo eixo coordenado horizontal é dado pela probabilidade de falso alarme, $P(S \mid n)$ e o eixo coordenado vertical é dado pela probabilidade de acerto, $P(S \mid s)$ [33], [8]. Consoante a mudança de critério de decisão do observador, cria-se um novo ponto no gráfico. Para um critério muito baixo, o ponto encontra-se situado nas coordenadas $(1, 1)$. À medida que o critério de decisão aumenta, as duas coordenadas diminuem e poder-se-à traçar uma curva contínua, até um critério de decisão muito elevado, que termina no ponto de coordenadas $(0, 0)$. O gráfico criado por este conjunto de pontos é designado por curva ROC "*sim-não*" [33], [8].

Green e Swets [33] demonstram que o declive da curva ROC em qualquer ponto é igual ao critério da razão de verosimilhanças que gera esse ponto. Considere-se o caso de variáveis contínuas, seja $f(e \mid s)$ a designação para a função densidade de probabilidade quando a hipótese s é verdadeira, e $f(e \mid n)$ a designação para a função densidade de probabilidade quando a hipótese n é verdadeira, então por definição:

$$l(e) = \frac{f(e|s)}{f(e|n)}$$

As coordenadas da curva *ROC* podem ser expressas como função de um critério c da seguinte forma:

$$P(S|s) = \int_c^{+\infty} f(e|s)\ de$$
$$P(S|n) = \int_c^{+\infty} f(e|n)\ de.$$

Diferenciando estas expressões em ordem ao limite inferior c, obtém-se:

$$\frac{d\ P(S|s)}{dc} = -f(c|s)$$
$$\frac{d\ P(S|n)}{dc} = -f(c|n).$$

Utilizando a regra, $dy/dx = (dy/dc)(dc/dx)$, o declive num ponto determinado pelo critério c, é [33]:

$$\left.\frac{d\ P(S|s)}{d\ P(S|n)}\right|_c = \frac{-f(c|s)}{-f(c|n)} = l(c).$$

Para o caso discreto, critérios sucessivos da razão de verosimilhanças determinam os declives entre pontos sucessivos da curva *ROC* [33]. Reciprocamente, se existem alguns pontos na curva ROC, e se o processo de decisão é baseado no critério da razão das verosimilhanças, então os valores do critério da razão de verosimilhanças poderá ser inferido a partir dos declives das linhas que unem pontos sucessivos.

Um resultado importante que advém desta relação, é que uma curva *ROC* baseada no critério da razão de verosimilhanças tem uma probabilidade de acerto que é uma função monótona crescente da probabilidade de falso alarme, e um declive que é monótono decrescente.

4.3 Procedimento de *"classificação"*

No procedimento de *"classificação"* utiliza-se o mesmo formato de apresentação do procedimento *"sim-não"*, como ilustra o esquema da figura 4.5. A sequência de ocorrência de acontecimentos físicos nos dois procedimentos é a mesma, o que difere é a natureza da resposta do observador. No procedimento *"sim-não"* a resposta é do tipo binário, no procedimento de *"classificação"* existe um maior número de respostas possíveis [33]. Por exemplo, respostas com cinco categorias cuja primeira categoria representa a certeza, s até à quinta categoria que representa quase a certeza de que s não está presente, consequentemente n está.

4.4 Procedimento de *"escolha forçada dupla" (2AFC)*

O procedimento de "*escolha forçada dupla*" *(2AFC)* difere do procedimento *"sim-não"* e do procedimento de *"classificação",* definidos anteriormente, no sentido de que dois intervalos de observação precedem a resposta.

Poder-se-à representar uma sequência típica de acontecimentos associados a este procedimento de detecção como exemplificado na figura 4.6, [33]: dois intervalos de observação são dados, o sinal ocorre sempre num deles e o observador é forçado a escolher um dos intervalos.

No procedimento de *2AFC* é usual utilizar-se uma notação ligeiramente diferente dos outros dois casos anteriores. Assim, designa-se por $\langle sn \rangle$ como o acontecimento que define a ocorrência de sinal no primeiro intervalo, mas não no segundo, de forma similar, $\langle ns \rangle$ define o acontecimento que indica a ocorrência de sinal no segundo intervalo, mas não no primeiro. Por exemplo,

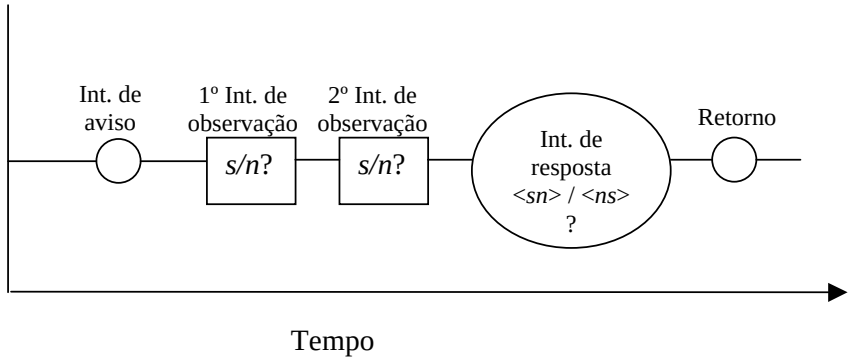


Figura 4.6: Acontecimentos numa experiência de procedimento "*escolha forçada dupla*" ($2AFC$).

na situação em que a ocorrência de sinal se verifica no primeiro intervalo temporal, isto é, $\langle sn \rangle$, um falso alarme corresponde aqui a decidir que o sinal ocorre no segundo intervalo temporal, e uma resposta correcta de que o sinal ocorre no primeiro intervalo, corresponde a um acerto.

Para proceder à comparação dos três procedimentos de decisão (*"sim-não*, *"classificação"* e *"2AFC"*), Green e Swets [33], assumem a existência de simetria na decisão do observador, no sentido de que não existe tendência na selecção de um intervalo relativamente a outro.

4.5 Teoria de detecção de sinal - relação entre o procedimento de escolha forçada dupla e as curvas ROC

No procedimento de escolha forçada dupla *(2AFC)* são considerados dois acontecimentos e_1 e e_2, que correspondem a cada intervalo de observações. Neste tipo de procedimento, o objectivo do observador consiste em decidir

se o primeiro é *sinal*, s e o segundo um *não sinal (ruído)*, n, ou o contrário.

Assume-se que o critério de decisão do observador é baseado no critério da *razão das verosimilhanças* [33] dado por:

$$l(e_i) = \frac{f(e_i \mid s)}{f(e_i \mid n)} \qquad (i = 1, 2). \tag{4.3}$$

No procedimento de escolha forçada dupla existem duas expressões para a razão de verosimilhanças, uma para cada acontecimento.

Assume-se que o observador escolhe o primeiro intervalo, se e só se, a razão de verosimilhanças associada a este intervalo é maior que a razão de verosimilhanças associada ao segundo [33].

Se a regra de decisão do observador é a selecção do intervalo que produza maior razão de verosimilhanças, ele estará correcto, se a razão de verosimilhanças associada à distribuição do *sinal+ruído* for maior do que a razão de verosimilhanças associada apenas à distribuição do *ruído.* Isto é, os dois intervalos do procedimento de escolha forçada dupla podem ser vistos como duas amostras aleatórias provenientes de duas distribuições estatísticas: uma designada por *sinal,* e outra por *ruído* (figura 4.7) [33].

Nesta situação pode-se considerar que o observador estará correcto, se a amostra proveniente da distribuição do *sinal* tiver uma maior razão de verosimilhanças do que a amostra proveniente do *ruído.* Suponha-se que o valor da razão de verosimilhanças retirado da distribuição do *sinal* é c; então o observador estará correcto se o valor da razão de verosimilhanças retirado da distribuição do *ruído* for menor que c.

Seja l_s a razão de verosimilhanças para distribuição do *sinal*, e l_n a razão de verosimilhanças para distribuição do *ruído*, o observador estará correcto se $l_s = c$ e $l_n < c$. Assim, se as duas amostras forem independentes, a probabilidade da ocorrência conjunta será o produto das duas probabilidades.

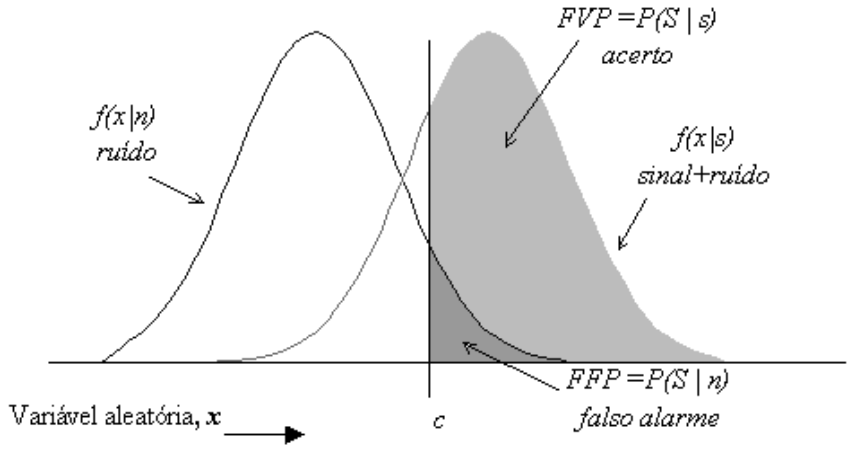


Figura 4.7: Distribuições hipotéticas para o *ruído* e para *sinal+ruído*.

Isto é, designando por $p_{2AFC}(C)$ a probabilidade de resposta correcta no procedimento de escolha forçada dupla:

$$p_{2AFC}(C) = P(l_s = c)\ P(l_n < c) \tag{4.4}$$

Assim, a probabilidade total do observador estar correcto será dada por [33]:

$$P_{2AFC}(C) = \int_{-\infty}^{+\infty} P(l_s = c)\ P(l_n < c)\ dc \tag{4.5}$$

Desde que l_s e l_n estejam distribuídas segundo $f(x \mid s)$ e $f(x \mid n)$, respectivamente, poder-se-à escrever a equação (4.5) como:

$$P_{2AFC}(C) = \int_{-\infty}^{+\infty} f(c \mid s) \quad \left[\int_{-\infty}^{c} f(x \mid n)\ dx\right]\ dc. \tag{4.6}$$

Agora a correspondência entre a percentagem de respostas correctas no procedimento de escolha forçada dupla e a classificação da curva ROC começa a emergir porque, como se irá demonstrar o lado direito da equação (4.6) envolve quantidades dadas pela curva ROC [33].

Se o critério de decisão do observador for c, a probabilidade de um falso alarme é dada por $P(l_n > c)$, isto é

$$P(l_n > c) = \int_c^{+\infty} f(x \mid n)\ dx = P_c\,(S \mid n) \tag{4.7}$$

ou

$$P(l_n < c) = \int_{-\infty}^{c} f(x \mid n)\ dx = 1 - P_c\,(S \mid n) \tag{4.8}$$

e

$$\frac{dP_c(S \mid s)}{dc} = \frac{d}{dc}\int_c^{+\infty} f(x \mid s)\ dx = -f(c \mid s). \tag{4.9}$$

Pode-se utilizar estas equações e substituir na equação (4.6). Note-se que a equação (4.9) dá a relação entre c e $P(S \mid s)$, assim os limites de integração podem ser determinados. Quando c é positivo e bastante elevado, então $P(S \mid s) = 0$ analogamente para largos valores negativos de c, $P(S \mid s) = 1$. Assim,

$$\begin{aligned} P_{2AFC}(C) &= -\int_1^0 dP_c\,(S \mid s)\ \ [1 - P_c\,(S \mid n)] \\ &= -\int_1^0 [1 - P_c\,(S \mid n)]\ \ dP_c\,(S \mid s) \end{aligned} \tag{4.10}$$

$$P_{2AFC}(C) = \int_0^1 [1 - P_c\,(S \mid n)]\ \ dP_c\,(S \mid s) \tag{4.11}$$

A ilustração desta demonstração em termos gráficos encontra-se na figura 4.8, que mostra que a percentagem de respostas correctas no procedimento de escolha forçada dupla é simplesmente a área abaixo da curva ROC no procedimento de resposta *"sim-não"*.

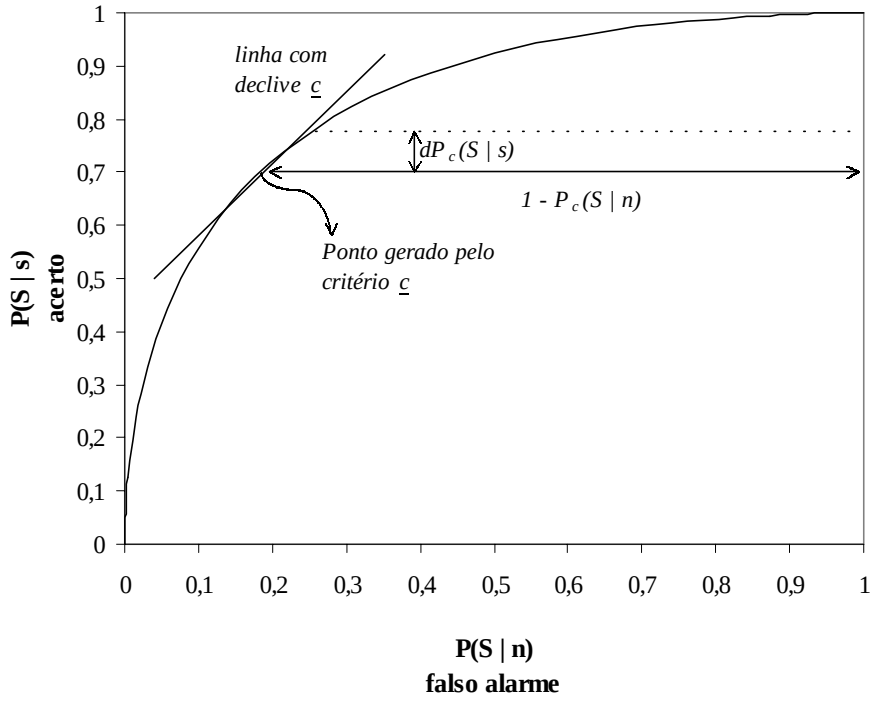


Figura 4.8: Relação entre a percentagem de respostas correctas na $2AFC$ e a área abaixo da curva ROC no procedimento *sim-não.*

Green e Swets [33] provaram ainda que a curva ROC para o procedimento de escolha forçada dupla é simétrica em relação à diagonal negativa, e os resultados seriam os mesmos se a escolha do observador recaísse sobre o segundo intervalo (e_2), isto é se a ordem de aparecimento do ruído fosse invertida.

A nota mais importante e que convém aqui salientar, é que todas as derivações são independentes da distribuição associada às hipóteses consideradas. Assim, provaram que a área abaixo da curva ROC é uma medida de probabilidade que é independente do tipo de distribuição associada, isto é, a área abaixo da curva ROC é uma estatística não paramétrica.

4.6 Análise de diagnóstico e a curva ROC

Charles E. Metz desenvolveu um conjunto de trabalhos sobre a aplicabilidade da análise *ROC* a sistemas de diagnóstico, nomeadamente no campo da imagem radiológica. Em [56] apresenta alguns princípios básicos da análise *ROC*, como o significado de *sensibilidade* e *especificidade* no desempenho dos testes e diagnóstico.

Define *sensibilidade* e *especificidade* como duas medidas de *precisão* de um teste de diagnóstico, dadas pelas fracções:

$$sensibilidade = \frac{\text{n}^o \text{ de decisões verdadeiras positivas}}{\text{n}^o \text{ de casos realmente positivos}}$$

$$especificidade = \frac{\text{n}^o \text{ de decisões verdadeiras negativas}}{\text{n}^o \text{ de casos realmente negativos}}.$$

Define também, *valor de corte*, como sendo um valor que pode ser seleccionado arbitrariamente de entre os valores possíveis para a variável de decisão, e acima do qual o paciente é classificado como *positivo* (teste de diagnóstico positivo, presença de doença), e abaixo do qual o paciente é classificado como *negativo* (teste de diagnóstico negativo, ausência de doença).

Assim, se existir alguma sobreposição entre a distribuição dos casos classificados como positivos e a distribuição dos casos classificados como negativos, e forçando o *valor de corte* a percorrer todos os valores possíveis da variável de decisão, podem-se obter vários pares de *fracções de verdadeiros positivos (sensibilidade)* e de *falsos positivos (1 - especificidade)*, que corresponderão, segundo Metz [56], aos eixos coordenados "y" e "x" de um gráfico que este designou por *curva ROC para o teste de diagnóstico.* Esta curva pode descrever as *características* de detecção associadas ao teste, e o observador pode *operar* em qualquer ponto da curva desde que seleccione o valor de corte

apropriado de decisão.

Para Metz [56] uma curva *ROC* convencional descreve os compromissos que podem ser tomados entre a FVP e a FFP, com a variação dos diferentes valores de corte ou critérios de decisão. Metz afirma que a análise *ROC* fornece uma descrição da detectabilidade da doença independentemente da prevalência desta e dos efeitos de escolha do critério de decisão.

Um outro aspecto, sobre o qual Metz se debruça [56], é a análise *custo/proveito* para um diagnóstico, tendo em conta as fracções definidas para a análise *ROC*.

O *custo médio* de um teste de diagnóstico, $\overline{C}$, é definido como [56]:

$$\begin{aligned} \overline{C} = C_0 &+ C_{VP} \;\; P(VP) + C_{VN} \;\; P(VN) \\ &+ C_{FP} \;\; P(FP) + C_{FN} \;\; P(FN) \end{aligned} \tag{4.12}$$

C_0: define o custo de realização do teste;

C_{VP}: define o custo médio das consequências médicas de uma decisão correctamente positiva (benefício);

$P(VP) = P(D^+) \;\; P(T^+ \mid D^+)$ com $P(D^+)$ a prevalência da doença em questão, e $P(T^+ \mid D^+)$ a proporção de indivíduos com teste positivo e que na realidade têm a doença;

C_{VN}: define o custo médio das consequências médicas de uma decisão correctamente negativa (benefício);

$P(VN) = P(D^-) \; P(T^- \mid D^-)$ com $P(D^-) = 1 - P(D^+)$, e $P(T^- \mid D^-)$ a proporção de indivíduos com teste negativo e que na realidade não têm a doença;

C_{FP}: define o custo médio das consequências médicas de uma decisão incorrectamente positiva;

$P(FP) = P(D^-)\ P(T^+ \mid D^-)$;

C_{FN}: define o custo médio das consequências médicas de uma decisão incorrectamente negativa;

$P(FN) = P(D^+)\ P(T^- \mid D^+)$;

Atendendo a que os benefícios poderão ser expressos como custos negativos, então a expressão (4.12) pode ser rearranjada, conduzindo a:

$$\begin{aligned}\overline{C} = &- \left\{[C_{FN} - C_{VP}]\ P(D^+)\right\}\ P(T^+ \mid D^+) \\ &+ \left\{[C_{FP} - C_{VN}]\ P(D^-)\right\}\ P(T^+ \mid D^-) \\ &+ \left\{C_0 + C_{VN}\ P(D^-) + C_{FN}\ P(D^+)\right\}\end{aligned} \tag{4.13}$$

Uma análise preliminar da expressão 4.13 revela que, independentemente dos custos médios das consequências de decisão, a média dos custos ($\overline{C}$) aumenta ou diminui consoante o custo de realização do teste (C_0). Assim, por exemplo, se um novo teste se revelar mais eficiente, isto é, fornecer melhores decisões em termos de diagnóstico, mas apresentar um custo de realização muito elevado, poder-se-à verificar um aumento do custo de diagnóstico.

Em 1983, Metz [60] desenvolveu um trabalho em que apresenta uma nova aproximação para testar diferenças significativas entre duas curvas *ROC* para dados correlacionados. Em 1986, Metz [58] apresenta um artigo onde efectua a análise estatística para dados *ROC* na avaliação de desempenho de diagnóstico, para os casos em que se tem duas amostras independentes ou duas amostras relacionadas. Neste artigo o autor descreve as propriedades estatísticas de um conjunto de dados *ROC* classificados (em imagem médica radiológica), procedimentos apropriados para o ajuste de uma curva *ROC* e ainda, testes que poderão ser utilizados para avaliar a significância estatística

da diferença aparente entre duas curvas *ROC*. Considera que os dados *ROC* classificados provêm de distribuições multinomiais que podem ser relacionadas com os parâmetros de um modelo subjacente em termos de teoria de detecção de sinal. Neste artigo, Metz [58] apresenta a curva *ROC* como sendo uma descrição empírica da capacidade de um sistema de diagnóstico para discriminar entre dois estados, onde cada ponto da curva representa um compromisso diferente entre as fracções já anteriormente definidas (FVP e a FFP), pela adopção de um *valor de corte de anormalidade* ou *nível crítico de confiança* diferente, no processo de decisão. Para procurar um ajuste para os dados, uma curva *ROC* pode ser descrita, assumindo que esta apresenta uma forma funcional particular com um ou mais parâmetros ajustáveis [58]. A forma funcional *binormal* para a curva *ROC* é utilizada em muitas situações práticas, revelando na sua maioria bons ajustes às curvas *ROC* empíricas [58]. Esta forma expressa as coordenadas da curva *ROC* através do par de expressões dadas pelas equações (2.11) e (2.12), como visto anteriormente.

Metz refere ainda que o *método de classes* (*rating method*) é muito utilizado na maior parte dos casos práticos [58]. Nesta aproximação é requerido ao observador que seleccione uma *classe* (*categoria* ou *confiança*) de entre algumas existentes. A utilização de k categorias fornece $k-1$ estimativas de pontos de operação na curva *ROC* convencional (para além dos pontos $(0,0)$ e $(1,1)$). Em diagnóstico de imagem médica são utilizadas, normalmente, cinco a seis categorias diferentes [58]. Os dados em classes, foram interpretados de acordo com um modelo desenvolvido por Metz [60], [58], que se passa a descrever.

Considerem-se I categorias, onde o observador pode definir $(I-1)$ valores de corte, c_i, no eixo da variável de decisão. A probabilidade da classe i é igual à probabilidade de que o resultado da variável decisão esteja entre c_{i-1}

e c_i, com $c_0 = -\infty$ e $c_I = +\infty$. Assim, para as imagens designadas por *realmente negativas*, a probabilidade de uma classe i é dada por:

$$p_i = \int_{c_i-1}^{c_i} f(x \mid n)\, dx \tag{4.14}$$

onde $f(x \mid n)$ é a função densidade de probabilidade na variável de decisão x para as imagens designadas por *realmente negativas.* De forma semelhante, pode-se definir:

$$\pi_i = \int_{c_i-1}^{c_i} f(x \mid a)\, dx \tag{4.15}$$

que corresponde à probabilidade de uma classe i, onde $f(x \mid a)$ é a função densidade de probabilidade na variável de decisão x para as imagens designadas por *realmente positivas.*

Se se considerar que a FFP representa a probabilidade de a variável de decisão x ter um valor maior ou igual que c_i para um ensaio *realmente negativo*, tem-se:

$$FFP(c_i) = \int_{c_i}^{+\infty} f(x \mid n)\, dx \tag{4.16}$$

$$= \sum_{j=i+1}^{I} p_j \tag{4.17}$$

considerando a equação (4.14), onde $1 \leq i \leq I-1$ e I é o número de categorias. De forma semelhante, a FVP associada ao valor de corte c_i pode ser dada por:

$$FVP(c_i) = \int_{c_i}^{+\infty} f(x \mid a)\, dx \tag{4.18}$$

$$= \sum_{j=i+1}^{I} \pi_j \tag{4.19}$$

Estas relações fornecem a base teórica para o cálculo das estimativas dos pontos da curva *ROC* para um conjunto de dados em classes [58].

Considere-se que os dados com I categorias são provenientes de M_n experiências independentes para os casos *realmente negativos*, e M_a experiências independentes para os casos *realmente positivos*. Os dados irão consistir em I números $k_i \ \ (1 \leq i \leq I)$ que representam o número de experiências *realmente negativas* na categoria i, e I números $l_i \ \ (1 \leq i \leq I)$ que representam o número de experiências *realmente positivas* na categoria i. Assim,

$$\sum_{i=1}^{I} k_i = M_n \tag{4.20}$$

e

$$\sum_{i=1}^{I} l_i = M_a. \tag{4.21}$$

Se as experiências forem independentes, então o conjunto de variáveis aleatórias $\{k_i : \ (1 \leq i \leq I)\}$ e $\{l_i : \ (1 \leq i \leq I)\}$ seguem uma distribuição multinomial com probabilidades de classe p_i e π_i, respectivamente.

A soma parcial:

$$K_{>i} = \sum_{j=i+1}^{I} k_j \tag{4.22}$$

representa o número de experiências *realmente negativas* para as quais a classe maior do que i foi obtida. Então para M_n experiências deste tipo, a soma parcial dada pela equação (4.22), segue uma distribuição binomial com valor esperado dado por $M_n \ FFP(c_i)$. Assim, para $(1 \leq i \leq I-1)$:

$$\widehat{FFP}(c_i) = \frac{K_{>i}}{M_n} \tag{4.23}$$

fornece uma estimativa não enviesada do valor de FFP associada ao iésimo observador com valor de corte c_i. De acordo com a estatística binomial, o desvio padrão desta estimativa é dado por:

$$\sigma_{\widehat{FFP}} = \sqrt{\frac{FFP(1-FFP)}{M_n}}. \tag{4.24}$$

De forma análoga, pode-se deduzir uma estimativa para FVP. Definindo a soma parcial [58],

$$L_{>i} = \sum_{j=i+1}^{I} l_j \tag{4.25}$$

que representa o número de experiências *realmente positivas* para as quais a classe maior do que i foi obtida. Para M_a experiências deste tipo, a estimativa não enviesada do valor de FVP associada ao iésimo observador com valor de corte c_i, pode ser dada por:

$$\widehat{FVP}(c_i) = \frac{L_{>i}}{M_a} \tag{4.26}$$

com desvio padrão:

$$\sigma_{\widehat{FVP}} = \sqrt{\frac{FVP(1-FVP)}{M_a}}. \tag{4.27}$$

Este procedimento descrito por Metz [58] pode ser posto em prática para obtenção de $(I-1)$ estimativas, para $1 \leq i \leq I-1$, de pares (FFP_i, FVP_i) na curva *ROC*. Estes $(I-1)$ pares coordenados, que vêm do canto superior direito para a esquerda no espaço unitário *ROC* com o decréscimo de i desde $(I-1)$ até 1, corresponde a $(I-1)$ valores de corte que o observador adopta na definição das I categorias de confiança que ele emprega.

Note-se que as $(I-1)$ estimativas dos pontos da curva *ROC* estão correlacionadas, porque os dados $\{k_j,\ l_j :\ (i+1 \leq j \leq I)\}$ utilizados para o cálculo

da *iésima* coordenada são incluídos no cálculo das coordenadas $(i-1)$, $(i-2)$, $(i-3)$, ...

Se não se considerar nenhuma forma funcional para a curva *ROC*, então as $(I-1)$ coordenadas dos pares *ROC* calculados directamente dos dados em classes, podem ser representados num plano unitário com barras de erro horizontais e verticais obtidas a partir das equações (4.24) e (4.27), e uma curva de ajuste pode ser traçada perto dos pontos e passando em $(0,0)$ e $(1,1)$.

Se considerar a forma funcional *binormal* para a curva *ROC*, as probabilidades p_i e π_i, anteriormente definidas, podem ser expressas na forma:

$$p_i = \Phi(c_i) - \Phi(c_{i-1}) \tag{4.28}$$

$$\pi_i = \Phi(b\ c_i - a) - \Phi(b\ c_{i-1} - a) \tag{4.29}$$

Desta forma, com os dados com I categorias, pode-se determinar os valores dos $(I+1)$ parâmetros ajustáveis $\{a,\ b,\ c_i :\ (1 \leq i \leq I-1)\}$ que produzam o melhor ajuste aos dados.

Metz, salienta, de entre os algoritmos para determinar as EMV, o desenvolvido por Dorfman e Alf [27], que utiliza o método de *scoring*, para resolução das equações não lineares resultantes da derivação em ordem aos parâmetros de interesse (ver anexo A). Neste trabalho [58] é desenvolvido sumariamente, a possibilidade deste sistemas de equações poder ser resolvido por um método iterativo do tipo de Newton-Raphson. Aponta que a diferença entre estes dois métodos, reside essencialmente no facto de que, no método de *scoring* as somas que envolvem as derivadas parciais de segunda ordem, referentes às probabilidades p_i e π_i, tendem para zero. Tal facto

torna este método computacionalmente mais estável. Por outro lado, este método acede automaticamente à precisão das estimativas que ele produz, produzindo as variâncias e covariâncias das estimativas dos parâmetros.

4.7 Relação entre o procedimento 2AFC e a análise de diagnóstico

Como já referido anteriormente, em diagnóstico interessa classificar os indivíduos como *normais* (baixos valores no eixo de decisão, x_N) e *anormais* (elevados valores no eixo de decisão, x_A). Nesta situação a área abaixo da curva *ROC* pode ser vista como uma medida de probabilidade de classificação correcta de um par (*normal, anormal).*

Fazendo o paralelismo entre o demonstrado por Green e Swets [33] para a *2AFC* e a análise de diagnóstico, c designa o *valor de corte*, θ corresponde à verdadeira área abaixo da curva *ROC* [37] ($P_{2AFC}(C)$), *normal* será a designação para *sinal* e *anormal* será a designação para *ruído.*

Assim, chega-se à conclusão

$$\theta = P_{2AFC}(C) = P(x_A > x_N)$$

em que $P(x_A > x_N)$ é a probabilidade de tomar uma decisão correcta em análise de diagnóstico.

4.8 Relação entre a área abaixo da curva ROC e a estatística de Wilcoxon-Mann-Whitney

Considere-se uma amostra de dimensão n_A para os indivíduos classificados como *anormais,* **A**, e outra de dimensão n_N para os indivíduos classificados como *normais,* **N**; o procedimento de teste consiste em fazer todas as $n_A\ n_N$ comparações possíveis entre os valores x_A da amostra n_A e os valores x_N da amostra n_N, graduando cada comparação de acordo com a regra,

$$T(x_A, x_N) = \begin{cases} 1 & \text{se } x_A > x_N \text{ } \textit{(concordante)} \\ \frac{1}{2} & \text{se } x_A = x_N \\ 0 & \text{se } x_A < x_N \text{ } \textit{(discordante)} \end{cases} \tag{4.30}$$

e fazendo a média dos T*s* para todas as $n_A\ n_N$ comparações, vem:

$$W = \frac{1}{n_A\ \ n_N}. \sum_{i=1}^{n_A} \sum_{j=1}^{n_N} T_{ij}(x_A, x_N) \tag{4.31}$$

que é uma estatística que não depende dos valores de x, mas apenas das graduações, designada como estatística de *Wilcoxon-Mann-Whitney* [37].

Como cada comparação é classificada por 1, $\frac{1}{2}$ ou 0, o valor médio de W estará entre 0 e 1, e reflecte, como não poderia deixar de ser, qual a proporção de x_A*́s* que são maiores que x_N.

Como nem todas as $n_A\ n_N$ comparações são independentes, inclui-las todas é mera conveniência, e o erro padrão de W tem em conta esta possível intercorrelação [37]. Assim, a probabilidade de atribuir uma classificação correcta é igual à média ponderada de todas as combinações de pares de classificações possíveis.

Seja A o acontecimento que designa a atribuição de uma classificação. Então $P(A)$, é dada por

$$P\,(\text{classificar correctamente}) = \frac{\text{n}^o \text{ de pares concordantes}}{\text{total de pares possíveis}}$$

Atendendo a que a distribuição dos x_A's se encontra à direita da distribuição dos valores de x_N, um par é classificado como concordante se e só se $x_A > x_N$. Se se considerar uma amostra de dimensão n_A para os indivíduos classificados como *anormais,* **A**, e outra de dimensão n_N para os indivíduos classificados como *normais,* **N**, definida a estatística T na equação (4.30), poder-se-à concluir que

$$\text{n }^o \text{ de pares concordantes} = \sum_{i=1}^{n_A}\sum_{j=1}^{n_N} T_{ij}(x_A, x_N).$$

Sendo $n_A\ n_N$ o número de comparações possíveis para cada par (*anormal, normal)*, o quociente destas duas quantidades:

$$P\,(\text{classificar correctamente}) = \frac{1}{n_A\ \ n_N}.\sum_{i=1}^{n_A}\sum_{j=1}^{n_N} T_{ij}(x_A, x_N) = W.$$

traduz o que se designou por estatística de Wilcoxon-Mann-Whitney, e também a proporção de pares correctamente classificados.

Assim, como provado por Green e Swets [33], $\theta = P_{2AFC}(C)$, e por outro lado em termos de análise de diagnóstico $P_{2AFC}(C) = P(x_A > x_N)$, decorrente da demonstração de Bamber [8], prova-se que

$$P\,(\text{classificar correctamente}) = W,$$

donde se pode concluir que

$$\theta = P_{2AFC}(C) = P(x_A > x_N)$$

Prova-se que a área abaixo da curva ROC é uma medida de probabilidade que é independente do tipo de distribuição associada, isto é, a área abaixo da

curva *ROC* pode ser quantificada através de uma estatística não paramétrica, nomeadamente a estatística de Wilcoxon-Mann-Whitney e, por conseguinte poder-se-à também determinar qual o erro padrão associado a esta medida, como consta no anexo B.

4.9 Distância perpendicular no plano *binormal*

Segundo Iverson (1992) [47], o desempenho na detecção, de um procedimento de resposta "*sim-não*" pode ser captado na forma de uma curva *ROC*, representando a fracção de verdadeiros positivos, "*acertos",* versus a fracção de falsos positivos, "*falsos alarmes"*. Assumindo a curva *ROC* como uma função estritamente crescente e que pode ser representada por um par (X_s, X_n) de variáveis aleatórias absolutamente contínuas, cada uma concentrada no contexto da recta real, de forma que, para cada critério numérico c, a probabilidade de um verdadeiro positivo *(acerto)*, p_{VP}, e a probabilidade de um falso positivo *(falso alarme)*, p_{FP}, têm funções "cauda", como ilustrado na figura 4.7 (página 69), dadas pelas expressões:

$$p_{VP} = P(X_s > c) \tag{4.32}$$

$$p_{FP} = P(X_n > c). \tag{4.33}$$

Assim X_n e X_s surgem como variáveis de decisão estatística. A aplicação da teoria de detecção de sinal a problemas específicos de detecção, envolve variáveis de decisão X_n e X_s gozando de propriedades especiais. Por exemplo, é frequente o caso em que X_n e X_s são do mesmo tipo, isto é, existem

quantidades $\sigma_s > 0$, $\sigma_n > 0$, μ_s e μ_n e uma variável aleatória X, independente da proveniência "s"ou ""n"tal que,

$$X_s = \sigma_s \, X + \mu_s \qquad \text{e} \qquad X_n = \sigma_n \, X + \mu_n \tag{4.34}$$

Considere-se o modelo padrão da teoria de detecção de sinal no qual X_n e X_s em (4.32) e (4.33) são gaussianas. As relações de (4.34) são automaticamente satisfeitas com a variável X distribuída como $Z \backsim N(0,1)$.

Sendo Φ a função de distribuição de Z, com $z_{VP} = \Phi^{-1}(p_{VP})$ e $z_{FP} = \Phi^{-1}(p_{FP})$, resulta de (4.32) e (4.33) que as variáveis z_{VP} e z_{FP} satisfazem a equação linear:

$$z_{VP} = m(z_{FP} + d') \tag{4.35}$$

com

$$m = \frac{\sigma_n}{\sigma_s} \qquad \text{e} \qquad d' = \frac{\mu_s - \mu_n}{\sigma_n} \tag{4.36}$$

e

$$z_{VP} = \frac{\mu_s - \mu_n}{\sigma_n} + \frac{\sigma_n}{\sigma_s} \, z_{FP}.$$

A ligação entre o procedimento "*sim-não*" e o da "*escolha-forçada*", como já referido anteriormente, é fornecida pelo "*teorema da área*" como descrito por Green e Swets [33], e traduzido pela expressão:

$$P_{2AFC} = P(X_s > X_n). \tag{4.37}$$

Quando as variáveis aleatórias X_s e X_n na equação (4.37) são gaussianas, obtém-se, em termos da notação referida em (4.34),

$$P_{2AFC} = P(\sigma_n \; Z - \sigma_s \; Z' < \mu_s - \mu_n)$$

onde Z e Z' são independentes e identicamente distribuídas segundo uma $N(0,1)$. Usando o facto de Z ser simétrico, e consequentemente

$$\sigma_n \; Z + \sigma_s \; Z' \sim \sqrt{\sigma_n^2 + \sigma_s^2} \; Z \tag{4.38}$$

determina-se:

$$P_{2AFC} = P\left(Z < \frac{\mu_s - \mu_n}{\sqrt{\sigma_n^2 + \sigma_s^2}} \right). \tag{4.39}$$

Escrevendo $z_c = \Phi^{-1}(P_{2AFC})$, as equações (4.35), (4.36) e (4.39) combinadas conduzem à equação linear nas variáveis z_{VP}, z_{FP} e z_c,

$$\sigma_s \; z_{VP} - \sigma_n \; z_{FP}{}' = \sqrt{\sigma_s^2 + \sigma_n^2} \; z_c. \tag{4.40}$$

É por vezes conveniente escrever a equação (4.40) na sua forma polar, fazendo,

$$\cos\theta = \frac{\sigma_s}{\sqrt{\sigma_s^2 + \sigma_n^2}} \qquad \text{e} \qquad \sin\theta = \frac{\sigma_n}{\sqrt{\sigma_s^2 + \sigma_n^2}} \qquad , \; 0 < \theta < \frac{\pi}{2}.$$

Assim,

$$\cos\theta \; z_{VP} - \sin\theta \; z_{FP}{}' = z_c \tag{4.41}$$

Esta última forma torna evidente que $z_c = \Phi^{-1}(P_{2AFC})$ corresponde à distância na perpendicular a partir da origem até à linha definida pela equação (4.35) cujo declive é $m = \tan\theta$.

Através da figura 4.9 e a partir da equação (4.35), chega-se a uma expressão para a área abaixo da curva *ROC* no plano *binormal*:

$$A_z = \Phi\left(\frac{a}{\sqrt{1+m^2}}\right)$$

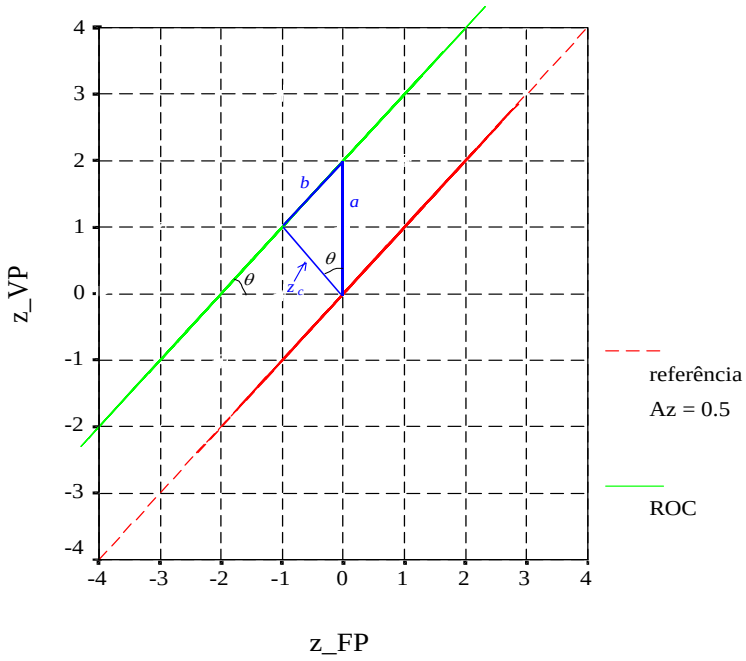


Figura 4.9: Relação da área abaixo da curva ROC com a distância na perpendicular no plano *binormal*.

Uma outra forma de demonstrar este resultado consiste em considerar as relações trigonométricas associadas a um ângulo θ. Assim, considerando o gráfico da figura 4.9, e das relações trigonométricas, sabe-se que para um determinado ângulo θ, tem-se:

$$\cos\theta = \frac{1}{\sqrt{1+(\tan\theta)^2}} \tag{4.42}$$

e para um triângulo rectângulo, as relações podem ser dadas por:

$$\cos\theta = \frac{\text{cateto adjacente}}{\text{hipotenusa}} = \frac{z_c}{a} \tag{4.43}$$

igualando as expressões (4.42) e (4.43), vem:

$$\frac{z_c}{a} = \frac{1}{\sqrt{1 + (\tan\theta)^2}}$$

Atendendo ainda a que $\tan\theta = b$ (ver figura 4.9), resulta:

$$z_c = \frac{a}{\sqrt{1+b^2}}.$$

Como z_c corresponde a um ponto do gráfico *binormal* cujos eixos coordenados são expressos em valores de *desvios normais,* este ponto representa um valor de $\Phi^{-1}(A)$, isto é, representa o *desvio normal* a que corresponde a probabilidade A. A expressão assim resultante para A_Z é da forma:

$$A_Z = \Phi\left(\frac{a}{\sqrt{1+b^2}}\right) \qquad c.q.d. \tag{4.44}$$

4.10 Comparação através de Curvas ROC

Uma das maiores virtualidades das curvas ROC consiste na possibilidade de comparar testes diferentes, como por exemplo, em diagnóstico médico. Em geral, constrói-se um teste de hipóteses efectuando o seguinte procedimento:

(1) escolha da hipótese nula que possa estar relacionada com os parâmetros da curva ROC;

(2) estimação dos parâmetros relevantes das duas curvas ROC, assim como as incertezas e correlações existentes nesses parâmetros;

(3) formação da estatística do teste que deverá seguir uma distribuição padrão se a hipótese nula for verdadeira;

(4) calculo do valor de prova (*valor-p*) de que um resultado da estatística de teste, pelo menos como extremo, poderá provir da distribuição assumida.

4.10.1 Amostras Independentes

Nos trabalhos desenvolvidos nesta área, Metz ([60], [58]) impõe como condições que os parâmetros podem ser estimados assumindo a forma funcional *binormal* para as curvas *ROC* utilizando como método de estimação o *método de scoring*.

Teste bivariado do Qui-Quadrado aos parâmetros

Assumindo a forma funcional *binormal* para as curvas ROC de dois sistemas de diagnóstico, x e y, estas podem ser especificadas pelos pares de parâmetros $(a_x,\ b_x)$ e $(a_y,\ b_y)$ que correspondem respectivamente, ao termo da ordenada na origem e declive na representação da curva ROC no plano *binormal*. A hipótese nula de que os dois conjuntos de dados em classes provêm de uma única *ROC* comum é equivalente a testar a hipótese:

$$H_0 :\ a_x = a_y\ \wedge\ b_x = b_y.$$

Se H_0 for verdadeira e as EMV, $\widehat{a}_x,\ \widehat{a}_y,\ \widehat{b}_x$ e $\widehat{b}_y$ seguem distribuição normal, então pode-se construir uma estatística de teste [58]

$$\mathbf{v} = \delta \mathbf{W}^{-1} \delta'$$

que segue uma distribuição do Qui-Quadrado com 2 graus de liberdade, e

$$\delta = \left(\widehat{a}_x - \widehat{a}_y,\ \widehat{b}_x - \widehat{b}_y\right)$$

$$\mathbf{W} = \begin{pmatrix} w_{11} & w_{12} \\ w_{21} & w_{22} \end{pmatrix}$$

com

$$w_{11} = var(\widehat{a}_x) + var(\widehat{a}_y) - 2\ cov\left(\widehat{a}_x, \widehat{a}_y\right) \tag{4.45}$$

$$w_{22} = var(\widehat{b}_x) + var(\widehat{b}_y) - 2\ cov\left(\widehat{b}_x, \widehat{b}_y\right) \tag{4.46}$$

$$\begin{aligned} w_{12} &= w_{21} \\ &= cov(\widehat{a}_x, \widehat{b}_x) + cov(\widehat{a}_y, \widehat{b}_y) - cov\left(\widehat{a}_x, \widehat{b}_y\right) - cov\left(\widehat{a}_y, \widehat{b}_x\right). \end{aligned} \tag{4.47}$$

Quando se trata de amostras independentes, os termos cruzados da matriz de covariâncias são nulos.

Teste à fracção de verdadeiros positivos, FVP

Existem algumas situações em que pode ter interesse verificar se dois sistemas de diagnóstico conduzem a curvas ROC com o mesmo valor de FVP num ponto particular de FFP_0 [58]. Então para dois sistemas de diagnóstico, x e y, a hipótese nula relevante é

$$H_0 :\ FVP_x\left(FFP_0\right) = FVP_y\left(FFP_0\right).$$

Quando duas curvas ROC se cruzam, esta hipótese nula pode ser verdadeira num ponto particular de FFP_0, quando o teste bivariado do Qui-Quadrado é falso [58].

Considerando a forma funcional *binormal* para cada uma das curvas ROC, então pode-se utilizar uma equação do tipo da referida na expressão (4.35) para testar a hipótese nula.

Assim:

$$a_x + b_x \ \Phi^{-1}\left(FFP_0\right) = a_y + b_y \ \Phi^{-1}\left(FFP_0\right) \tag{4.48}$$

ou então, com $c_0 = \Phi^{-1}\left(1 - FFP_0\right) = -\Phi^{-1}\left(FFP_0\right)$

$$\left(b_x - b_y\right) \ c_0 - \left(a_x - a_y\right) = 0 \tag{4.49}$$

Se H_0 for verdadeira e as EMV, $\widehat{a}_x$, $\widehat{a}_y$, $\widehat{b}_x$ e $\widehat{b}_y$ seguem distribuição normal, então pode-se construir uma estatística de teste [58]

$$v = \left(\widehat{b}_x - \widehat{b}_y\right) \ c_0 - \left(\widehat{a}_x - \widehat{a}_y\right)$$

cuja distribuição é Normal com média zero e desvio padrão

$$\sigma_v = \sqrt{w_{11} - 2 \ c_0 w_{12} + b_0^2 \ w_{22}}$$

onde os w_{ij}'s são dados pelas equações (4.45)-(4.47). Da mesma forma, tratando-se de amostras independentes, os termos cruzados da matriz de covariâncias são nulos.

Teste à área abaixo da curva ROC, A_Z

Aqui a hipótese nula relevante assume que os dois conjuntos de dados, em classes, provêm de curvas ROC com igual área abaixo desta.

$$H_0 : \ A_{Z_x} = A_{Z_y}$$

Na situação em que duas curvas ROC se cruzam, a hipótese nula do teste da área pode ser verdadeira, quando a do teste bivariado é falsa e a do teste à FVP é falsa, excepto num ponto único de FFP_0 [58].

A equação (4.44) expressa o índice A_Z em termos de dois parâmetros da curva *ROC binormal*. Para um número de experiências elevado, as incertezas relativas nas estimativas dos parâmetros das duas curvas *ROC* ($\widehat{a}_x$, $\widehat{a}_y$, $\widehat{b}_x$ e $\widehat{b}_y$) tornam-se pequenas, e estas estimativas aproximam-se de uma Normal. Assim, se H_0 for verdadeira, a diferença entre os índices A_Z para dois sistemas de diagnóstico, x e y:

$$v = \Phi\left(\frac{\widehat{a}_x}{\sqrt{1+\widehat{b}_x^2}}\right) - \Phi\left(\frac{\widehat{a}_y}{\sqrt{1+\widehat{b}_y^2}}\right) \tag{4.50}$$

segue aproximadamente distribuição Normal com média zero e variância [58]

$$\sigma_v^2 = \sum_{i=1}^{4}\sum_{j=1}^{4}\left(\frac{\partial v}{\partial \theta_i}\right)\left(\frac{\partial v}{\partial \theta_j}\right)\ cov\left(\widehat{\theta}_i, \widehat{\theta}_j\right) \tag{4.51}$$

onde $\{\theta_i : i = 1, 2, 3, 4\} = \{a_x, a_y, b_x, b_y\}$ representa o conjunto dos quatro parâmetros das duas curvas *ROC*. Tratando-se de amostras independentes, os termos cruzados da matriz de covariâncias na equação (4.51) são nulos, e os restantes termos poderão ser estimados pelo *método de scoring.*

Teste à área abaixo da curva ROC, A − Abordagem não paramétrica

Um outro método para testar se as diferenças entre duas áreas abaixo das curvas *ROC* provenientes de amostras independentes são significativas, consiste na utilização da razão crítica z, definida por Hanley e McNeil [38]:

$$z = \frac{A_1 - A_2}{\sqrt{SE_1^2 + SE_2^2}} \backsim N(0,1).$$

As áreas abaixo das curvas *ROC* para cada uma das modalidades a comparar (A_1 e A_2) e os erros padrão respectivos (SE_1 e SE_2), são obtidos através da aproximação à estatística de Wilcoxon-Mann-Whitney. Quando

os valores da área abaixo da curva *ROC* são superiores a 0.5, os erros padrão associados ás áreas, podem ser obtidos através da expressão [37]

$$SE(A) = \sqrt{\frac{A\,(1-A) + (n_A - 1)\,(Q_1 - A^2) + (n_N - 1)\,(Q_2 - A^2)}{n_A\ n_N}} \quad (4.52)$$

em que Q_1 corresponde à probabilidade de duas imagens anormais, aleatoriamente escolhidas serem classificadas com maior suspeição do que uma imagem normal aleatoriamente escolhida, e Q_2 corresponde à probabilidade de uma imagem anormal, aleatoriamente escolhida ser classificada com maior suspeição do que duas imagens normais aleatoriamente escolhidas.

Hanley e McNeil [37] provaram que sob a assumpção do modelo exponencial negativo (modelo que conduziu a valores de erros padrão mais conservativos quando comparado com outros modelos, como o Gaussiano ou Gama), Q_1 e Q_2 podem ser expressos como uma função do índice área abaixo da curva ROC, isto é,

$$Q_1 = \frac{A}{2 - A}$$

$$Q_2 = \frac{2\ A^2}{1 + A}.$$

A substituição destas expressões na equação (4.52) conduz ao valor de erro padrão esperado para qualquer valor de A.

4.10.2 Amostras correlacionadas

Para detectar correctamente uma diferença significativa entre curvas *ROC* medidas no mesmo paciente ou na mesma imagem, o efeito da covariância na variância da diferença deverá ser estimada e incorporada no teste. Para

conjuntos de dados correlacionados, os termos das covariâncias nas equações (4.45), (4.47) e (4.51) são normalmente diferentes de zero, necessitando, por isso, de ser estimados por um método.

A aproximação desenvolvida por Metz [60] é baseada na generalização do modelo *binormal*, que este designou por *modelo binormal bivariado*.

Considera-se neste modelo, duas variáveis de decisão x e y correlacionadas, provenientes de uma de duas funções de densidade de probabilidade conjuntas Normais, $f(x, y \mid n)$ e $f(x, y \mid a)$. Cada uma destas densidades tem médias e desvios padrões diferentes nas direcções de x e y, e cada uma é caracterizada por diferentes coeficientes de correlação, r_n e r_a, como se pode ver no exemplo da figura 4.10.

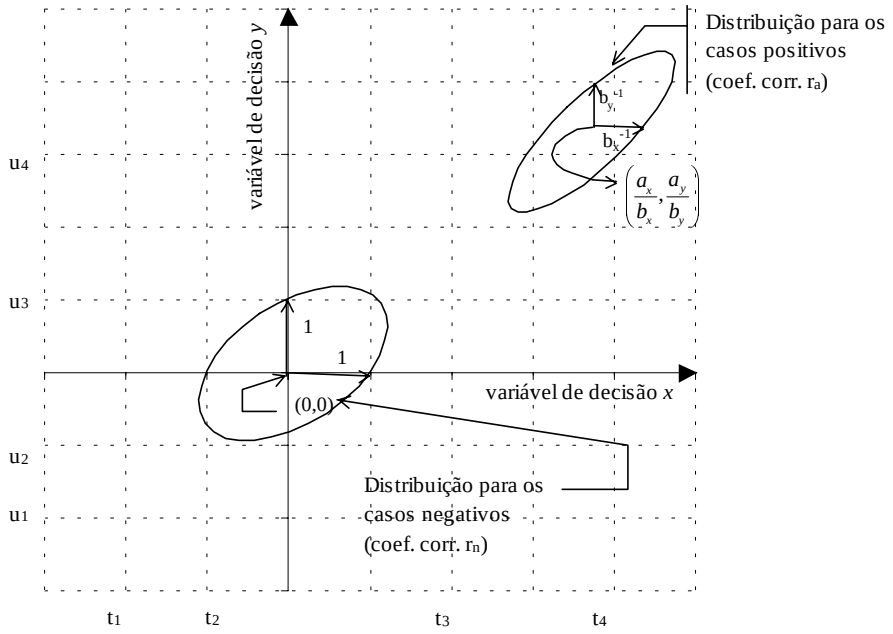


Figura 4.10: Exemplo esquemático do modelo bivariado.

Considerem-se as seguintes notações, introduzidas por Metz [60]:

p_{ij}: probabilidade de um par de categorias i e j das imagens

consideradas negativas, n;

π_{ij}: probabilidade de um par de categorias i e j das imagens consideradas positivas, a;

t_{i-1} e t_i: barreiras na variável de decisão x;

u_{j-1} e u_j: barreiras na variável de decisão y;

$(a_x,\ b_x)$: parâmetros da curva ROC quando a observação em x é tida individualmente;

$(a_y,\ b_y)$: parâmetros da curva ROC quando a observação em y é tida individualmente.

Assim,

$$\begin{aligned} p_{ij} &= L(t_i, u_j, r_n) + L(t_{i-1}, u_{j-1}, r_n) \\ &\quad - L(t_{i-1}, u_j, r_n) - L(t_i, u_{j-1}, r_n) \end{aligned} \tag{4.53}$$

onde $L(x, y, r)$ é a função de distribuição acumulada para a normal bivariada:

$$L(x, y, r) = \int_x^{+\infty} dv \int_y^{+\infty} g(v, w, r)\ dw, \qquad \text{com}$$

$$g(v, w, r) = \frac{1}{2\pi\sqrt{1 - r^2}} \exp\left[-\frac{x^2 - 2rxy + y^2}{2\,(1 - r^2)}\right]$$

a função densidade de probabilidade da Normal bivariada padrão.

De forma análoga, para uma experiência *realmente positiva*, define-se:

$$\begin{aligned} \pi_{ij} &= L(b_x t_i - a_x, b_y u_j - a_y, r_a) + L(b_x t_{i-1} - a_x, b_y u_{j-1} - a_y, r_a) \\ &\quad - L(b_x t_{i-1} - a_x, b_y u_j - a_y, r_a) - L(b_x t_i - a_x, b_y u_{j-1} - a_y, r_a) \end{aligned} \tag{4.54}$$

Segundo Metz [58] o *método de scoring* poderá ser utilizado para determinar as EMV para os parâmetros do modelo *binormal bivariado* para os dados em classes e correlacionados. Com estas estimativas, e com as estimativas dos termos cruzados da matriz de covariâncias, para as curvas *ROC* em estudo, pode-se aplicar qualquer um dos testes descritos para o caso de duas amostras independentes, tendo em consideração os valores das covariâncias.

Qualquer um dos três testes mencionados, só é exacto no limite dos grandes números, mas, no entanto, apresentam um bom desempenho para amostras com 50 casos de cada tipo (*negativo* e *positivo*) [58].

Metz [58] refere que uma aproximação útil para testar as diferenças para dados correlacionados foi a de Hanley e McNeil [55], que emprega a estatística de Wilcoxon-Man-Whitney para dados correlacionados para o teste ao índice área abaixo da curva *ROC*.

Teste à área abaixo da curva ROC, A − Abordagem não paramétrica

A razão crítica z permite testar se as diferenças entre duas áreas abaixo das curvas *ROC,* provenientes do mesmo conjunto de dados, são aleatórias ou significativas. Esta razão é definida como [38]:

$$z = \frac{A_1 - A_2}{\sqrt{SE_1^2 + SE_2^2 - 2\ r\ SE_1\ SE_2}}$$

onde A_1 e SE_1 e A_2 e SE_2 correspondem às áreas observadas e erros padrão estimados da curva *ROC* para as modalidades 1 e 2, respectivamente; r, representa a correlação estimada entre A_1 e A_2.

Esta quantidade z reporta-se às tabelas da distribuição Normal padrão, e valores de z acima de um determinado valor evidenciam, estatisticamente, que as verdadeiras "*áreas ROC*" são diferentes. A importância da introdução do termo $2\ r\ SE_1\ SE_2$ no denominador da expressão é devido ao facto de

os dados estarem correlacionados, porque foram recolhidos sobre a mesma amostra, e a ausência deste termo implicaria um denominador de maior valor e, consequentemente, o valor de z mais pequeno o que, eventualmente, reduziria a possibilidade de detectar diferenças significativas entre as duas modalidades [38].

Cálculo do coeficiente de correlação entre áreas O método sugerido por Hanley e McNeil em [38] utiliza uma tabela para determinação do coeficiente de correlação entre áreas, r. Calculam-se dois coeficientes de correlação intermédios, que são depois convertidos à correlação entre A_1 e A_2 através da técnica sugerida em [38]. Assim, determina-se o coeficiente de correlação, r_N, para as classificações dadas para os pacientes normais e o coeficiente de correlação, r_A, para as classificações dadas para os pacientes anormais para as duas modalidades. Cada um destes coeficientes pode ser calculado pelas formas tradicionais utilizando, quer o método de cálculo do produto dos momentos para a correlação de Pearson, quer o método do tau de Kendall. Como as curvas *ROC* em medicina são normalmente obtidas a partir de dados numa escala ordinal, utiliza-se o tau de Kendall para calcular r_N e r_A.

O coeficiente de correlação médio, $(r_N+r_A)/2$, e a área média, $(A_1+A_2)/2$, vão constituir as entradas numa tabela construída por Hanley e McNeil [38], a partir da qual se retira o valor de r.

Os coeficientes de correlação entre áreas podem também ser determinados através do método sugerido por DeLong e DeLong em [22]. Este método utiliza uma aproximação não paramétrica ao cálculo da área abaixo de curvas *ROC*, para conjuntos de dados correlacionados, utilizando a teoria das estatísticas *U-generalizadas* para estimação da matriz de covariâncias, quando

se comparam duas ou mais curvas *ROC*. Segundo este método, admitindo que se tem m indivíduos que apresentam na realidade a doença e n indivíduos que não têm a doença, a matriz de covariâncias estimada para o vector de parâmetros (área abaixo da curva *ROC*) $\widehat{\theta} = \left(\widehat{\theta}^1, \widehat{\theta}^2, ..., \widehat{\theta}^k\right)$, em que k representa o número de modalidades a comparar, é tal que:

$$\mathbf{S} = \frac{1}{m}\mathbf{S}_{10} + \frac{1}{n}\mathbf{S}_{01}.$$

Seja $X_i,\ i = 1, 2, ..., m$ e $Y_j,\ j = 1, 2, ..., n$ os valores das variáveis nos quais o teste de diagnóstico é baseado, e supondo que valores elevados da variável teste estão associados à presença de doença, as matrizes $\mathbf{S}_{10}$ e $\mathbf{S}_{01}$ com dimensão $k \times k$ são definidas, respectivamente, para o (r, s) elemento, pelas expressões das equações (4.55) e (4.56):

$$s_{10}^{r,s} = \frac{1}{m-1}\sum_{i=1}^{m}\left[V_{10}^{r}(X_i) - \widehat{\theta}^r\right]\left[V_{10}^{s}(X_i) - \widehat{\theta}^s\right] \tag{4.55}$$

$$s_{01}^{r,s} = \frac{1}{n-1}\sum_{j=1}^{n}\left[V_{01}^{r}(Y_j) - \widehat{\theta}^r\right]\left[V_{01}^{s}(Y_j) - \widehat{\theta}^s\right] \tag{4.56}$$

V_{10}^r e V_{01}^r, representam as componentes em X e Y, respectivamente, para a r-ésima estatística $\widehat{\theta}^r$, definidas por:

$$V_{10}^{r} = \frac{1}{n}\sum_{j=1}^{n}\psi\left(X_i^r, Y_j^r\right) \qquad (\ i = 1, 2, ..., m) \tag{4.57}$$

$$V_{01}^{r} = \frac{1}{m}\sum_{i=1}^{m}\psi\left(X_i^r, Y_j^r\right) \qquad (j = 1, 2, ..., n) \tag{4.58}$$

com $\psi(X, Y)$ definida através da expressão da equação (4.59):

$$\psi(X,Y) = \begin{cases} 1 & Y < X \\ \frac{1}{2} & Y = X \\ 0 & Y > X \end{cases} \tag{4.59}$$

De salientar que a média desta função $\psi(X,Y)$, conduz à estimativa da estatística de Mann-Whitney, que corresponde a um estimador da área abaixo da curva *ROC* $(\widehat{\theta})$, como referido anteriormente.

Capítulo 5

Análise da curva ROC

5.1 Relação entre as funções densidade de probabilidade associadas aos dados e a forma da curva ROC

Com este estudo pretende–se verificar como é que algumas hipóteses sobre as distribuições associadas à variável de decisão podem afectar a forma da curva ROC.

Com base na hipótese da Normalidade, e através de estudos de simulação procurou-se numa primeira abordagem verificar qual a variação da forma da curva ROC em função do parâmetro de localização e/ou de escala para a função densidade de probabilidade dos casos designados por anormais (maiores valores da variável de decisão).

Numa segunda abordagem, criaram-se as seguintes hipóteses para os casos denominados normais e para os casos denominados anormais:

(i) duas distribuições Normais;

(ii) duas distribuições Logísticas e de igual variância;

(iii) duas distribuições Exponenciais negativas com diferentes parâmetros de escala θ;

(iv) duas distribuições Uniformes num intervalo (a, b).

Para a visualização da curva ROC, utilizou-se a representação desta no plano ROC unitário e no plano *binormal*.

Autores como Swets [80], afirmam que a forma das funções densidade de probabilidade na variável de decisão determina a forma da curva ROC. Procurou-se assim, através de métodos gráficos (no plano unitário e no plano *binormal*) e analíticos, validar ou não esta afirmação.

Para o estudo sobre a forma da curva ROC consideraram-se as seguintes condições:

- a variável de decisão é contínua;
- os casos designados por normais correspondem a valores menores da variável de decisão, e os casos designados por anormais correspondem aos maiores valores da variável de decisão;
- existe sobreposição entre as funções densidade de probabilidade, isto é, existe uma área de sobreposição entre as duas funções densidade de probabilidade associadas aos casos normais e anormais.

Com base nestas condições realizaram-se os estudos de simulação descritos nas secções seguintes.

5.1.1 Funções densidade de probabilidade Normais

A hipótese da Normalidade é a mais utilizada no desenvolvimento da teoria clássica de detecção de sinal, pelo que foi considerada em primeiro lugar. Admita-se que a variável x tem distribuição Normal, com média μ_N e variância σ_N^2 sob h_0, e com média μ_A e variância σ_A^2 sob h_1, pelo que as funções densidade de probabilidade respectivas são:

$$f(x|h_0) = \frac{1}{\sigma_N\sqrt{2\pi}} \exp\left[-\frac{(x-\mu_N)^2}{2\sigma_N^2}\right] \qquad \begin{array}{r} -\infty < x < +\infty \\ -\infty < \mu_N < +\infty, \quad \sigma_N > 0 \end{array} \tag{5.1}$$

e

$$f(x|h_1) = \frac{1}{\sigma_A\sqrt{2\pi}} \exp\left[-\frac{(x-\mu_A)^2}{2\sigma_A^2}\right] \qquad \begin{array}{r} -\infty < x < +\infty \\ -\infty < \mu_A < +\infty, \quad \sigma_A > 0 \end{array} \tag{5.2}$$

As coordenadas da curva ROC determinadas para as hipóteses definidas anteriormente, e para um dado valor de corte c, serão dadas por:

$$FVP = P(H_1 \mid h_1) = \int_c^{+\infty} f(x \mid h_1)\, dx = 1 - \Phi\left(\frac{c-\mu_A}{\sigma_A}\right) \tag{5.3}$$

$$FFP = P(H_1 \mid h_0) = \int_c^{+\infty} f(x \mid h_0)\, dx = 1 - \Phi\left(\frac{c-\mu_N}{\sigma_N}\right) \tag{5.4}$$

onde Φ representa a função distribuição acumulada da Normal padrão.

As equações (5.3) e (5.4) contêm inicialmente quatro parâmetros de interesse, relativos às duas distribuições Normais mas, de facto, apenas dois

desses parâmetros são relevantes para a análise. Considerando a seguinte transformação de variável,

$$y = \frac{1}{\sigma_N} \ (x - \mu_N) \tag{5.5}$$

faz com que a distribuição de x segundo h_0 tenha uma média zero e desvio padrão unitário e a distribuição, segundo h_1, terá uma média de $(\mu_A - \mu_N)/\sigma_N$ e um desvio padrão de σ_A/σ_N. Com esta transformação, a distância entre as duas médias $(\mu_A - \mu_N)$ e a razão dos desvios padrão σ_A/σ_N são os parâmetros de interesse [33].

Reescrevendo as expressões (5.3) e (5.4), tendo em consideração a transformação linear referida em (5.5), obtém-se:

$$FVP = \Phi(d' \ b - b \ q) \tag{5.6}$$

$$FFP = \Phi\left(-q\right) \tag{5.7}$$

Este par de equações fornece uma forma funcional para a curva ROC, em função dos parâmetros "d'" e "b" e para um dado valor de corte c com $q = (c - \mu_N)/\sigma_N$. Tendo em conta a transformação linear sugerida, os parâmetros "d'" e "b" terão as seguintes expressões:

$$d' = \frac{(\mu_A - \mu_N)}{\sigma_N} \tag{5.8}$$

$$b = \frac{\sigma_N}{\sigma_A} \tag{5.9}$$

As equações (5.6) e (5.7) mostram que a curva ROC pode ser dada explicitamente pela forma:

$$FVP = \Phi\left(d'\ b + b\ \Phi^{-1}\left(FFP\right)\right) \tag{5.10}$$

Nos eixos coordenados do papel de probabilidades binormal, a curva ROC é dada por:

$$\Phi^{-1}\left(FVP\right) = d'\ b + b\ \Phi^{-1}\left(FFP\right) \tag{5.11}$$

onde $\Phi^{-1}(FVP)$ e $\Phi^{-1}(FFP)$ representam os eixos coordenados correspondentes às probabilidades FVP (sensibilidade) e FFP (1-especificidade), "$d'b = a$" é a ordenada na origem e "b" o declive da recta representada no plano *binormal*. Consequentemente, a expressão para o cálculo do valor da área abaixo da curva ROC será dada por:

$$A_z = \Phi\left(\frac{a}{\sqrt{1+b^2}}\right) \tag{5.12}$$

Johnson, Kotz e Balakrishnan [49] apresentam várias aproximações a Φ^{-1}, que podem ser usadas para examinar analiticamente as curvas no plano *binormal*, como:

$$\Phi^{-1}\left(p\right) \approx -5.5310\left\{\left(\tfrac{1-p}{p}\right)^{0.1193} - 1\right\}, \quad p > \tfrac{1}{2}$$

$$\Phi^{-1}\left(p\right) \approx -0.4115\left\{\tfrac{1-p}{p} + \ln\left(\tfrac{1-p}{p}\right) - 1\right\}, \quad p \geq \tfrac{1}{2}$$

$$\Phi^{-1}(p) \approx -a' \ln\left(\frac{1-p}{p}\right) + b', \quad p \geq \frac{1}{2}$$

onde a' e b' deverão satisfazer a condição

$$b' = \sqrt{1 - 1.3682\ a'} - 1.3862\ a'$$

No entanto estas expressões podem não produzir linhas rectas no caso Normal devido aos erros associados às aproximações.

Funções densidade de probabilidade Normais com igual variância

As funções densidade de probabilidade Normais de igual variância foram as primeiras a serem consideradas por Thurstone [80], como referido anteriormente. Nesta situação o índice de discriminação mais utilizado é:

$$d' = \frac{\mu_A - \mu_N}{\sigma}$$

que expressa a diferença entre as médias das duas funções densidade de probabilidade em termos de desvio padrão.

As coordenadas da curva ROC, para estas hipóteses, e para um dado *valor de corte c* são dadas pelas expressões das equações (5.3) e (5.4). A equação (5.11), dá a forma da curva ROC para a situação em estudo.

Procurando ilustrar a situação descrita, geraram-se amostras aleatórias com valores de $\mu_N = 50$ e $\sigma_N = \sigma_A = \sigma = 5$, e $\mu_A = 55$, $\mu_A = 60$ e $\mu_A = 70$, e obtiveram-se os resultados apresentados na figura 5.1 em termos de representação no plano ROC unitário, e na figura 5.2 em termos de plano *binormal*. Os valores escolhidos para os parâmetros não são representativos de nenhuma situação em particular, dado que o objectivo deste estudo é averiguar o efeito da diferença entre as médias e a razão dos desvios, σ_A/σ_N.

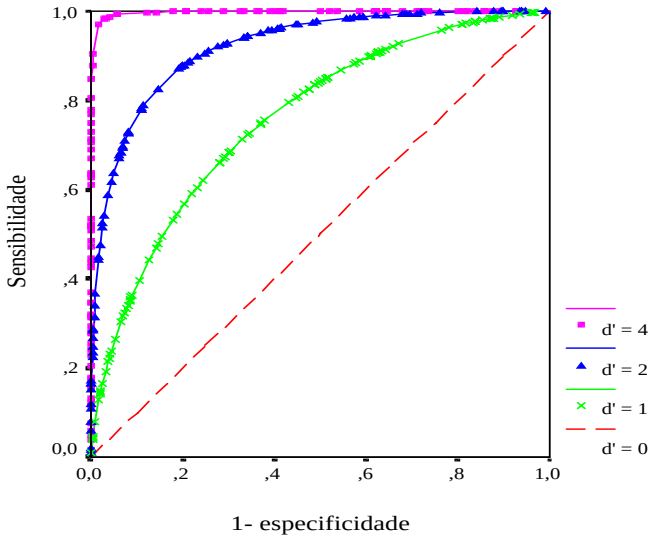


Figura 5.1: Representação das curvas ROC para distribuições Normais de igual variância no plano ROC.

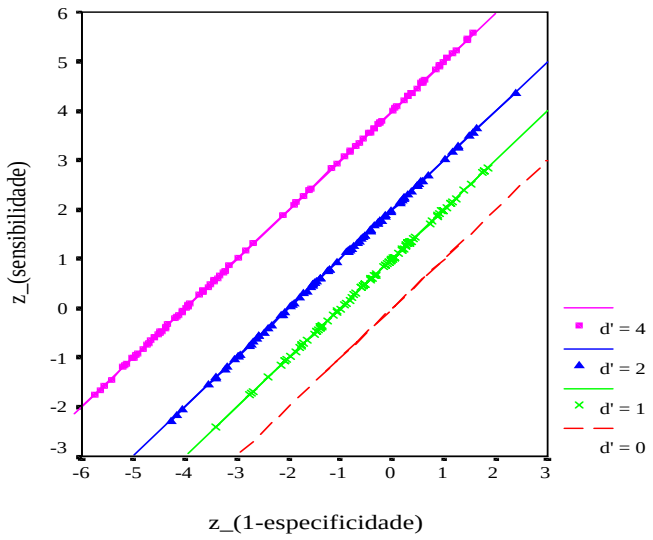


Figura 5.2: Representação das curvas ROC para distribuições Normais de igual variância no plano *binormal*.

A partir da análise da figura 5.1 verifica-se que à medida que d' aumenta, o poder discriminante também aumenta (curva mais chegada ao canto superior esquerdo a que corresponde $d' = 4$), e consequentemente, obtém-se um maior valor de área abaixo da curva ROC, mantendo-se no entanto a forma da curva.

Da análise da figura 5.2, é importante salientar que a representação no plano *binormal* de cada uma destas curvas é uma recta bem definida, podendo-se assim, determinar os valores dos respectivos declives e das ordenadas na origem, através da expressão definida na equação (5.11).

Funções densidade de probabilidade Normais com $\sigma_N^2 \neq \sigma_A^2$

No caso das função densidade de probabilidade para os casos designados por anormais e para os casos designados por normais terem variâncias diferentes, foram consideradas as seguintes situações:

a) $\frac{\sigma_A}{\sigma_N} > 1$;
b) $\frac{\sigma_A}{\sigma_N} < 1$

Hipoteticamente observar-se-ia uma situação semelhante à ilustrada nas figuras 5.3 e 5.4, respectivamente.

Experimentalmente considerou-se, $\mu_N = 50$, $\mu_A = 60$ e $\frac{\sigma_A}{\sigma_N} = 4$. Os resultados obtidos encontram-se nas figuras 5.5 e 5.6.

Para a segunda situação $\mu_N = 50, \mu_A = 60$ e $\frac{\sigma_A}{\sigma_N} = 0.25$. Os resultados obtidos encontram-se nas figuras 5.7 e 5.8.

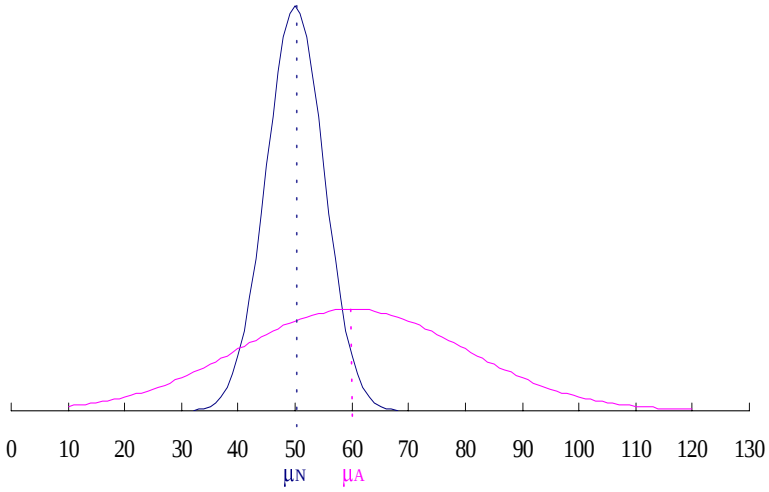


Figura 5.3: Sobreposição de 2 distribuições para o caso a).

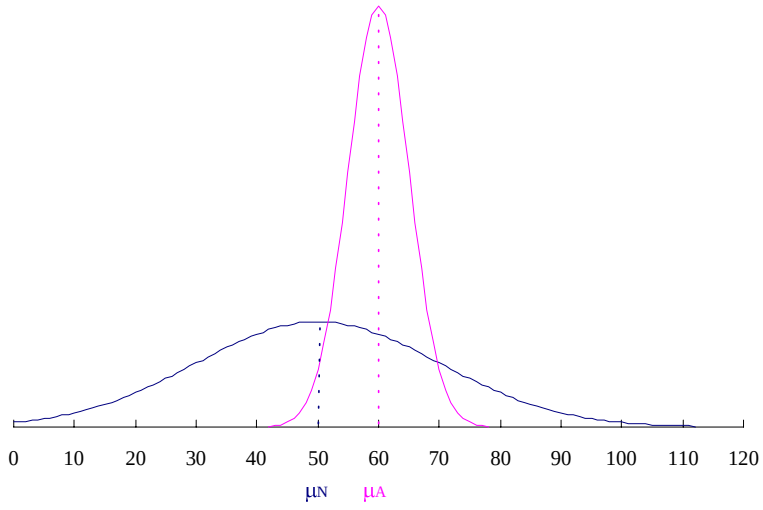


Figura 5.4: Sobreposição de 2 distribuições para o caso b).

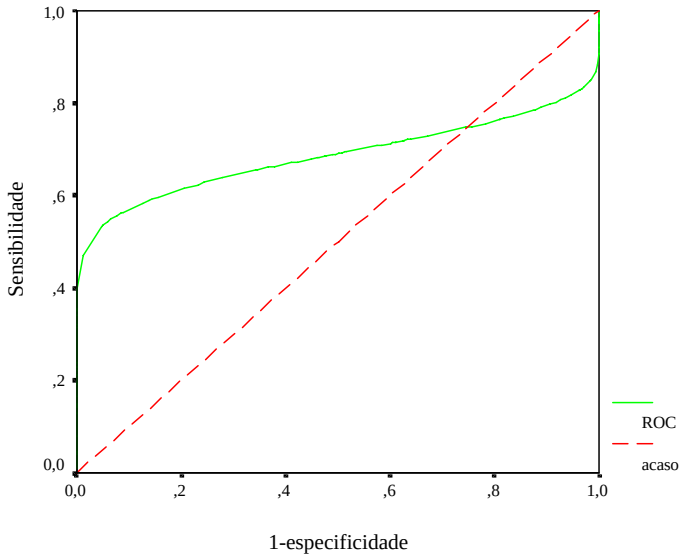


Figura 5.5: Representação da curva ROC para a situação descrita em a), no plano ROC.

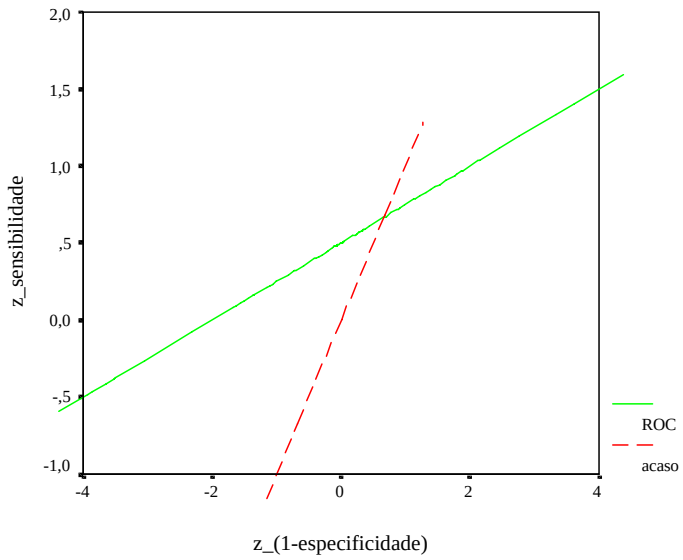


Figura 5.6: Representação da curva ROC para a situação descrita em a), no plano *binormal*.

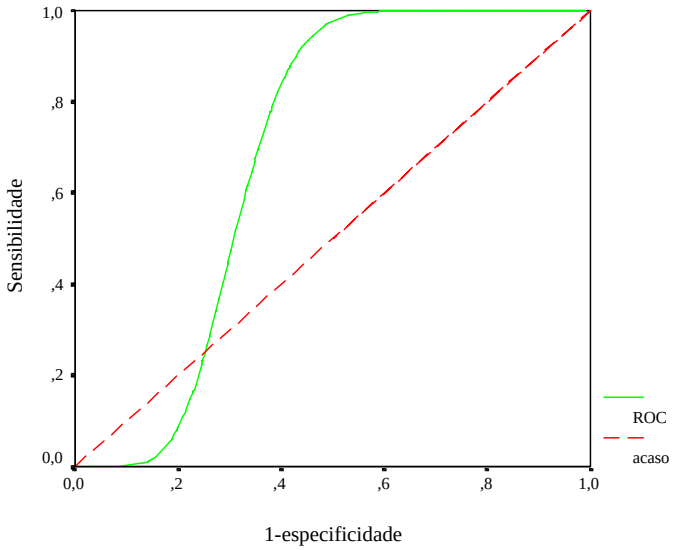


Figura 5.7: Representação da curva ROC para a situação descrita em b), no plano ROC.

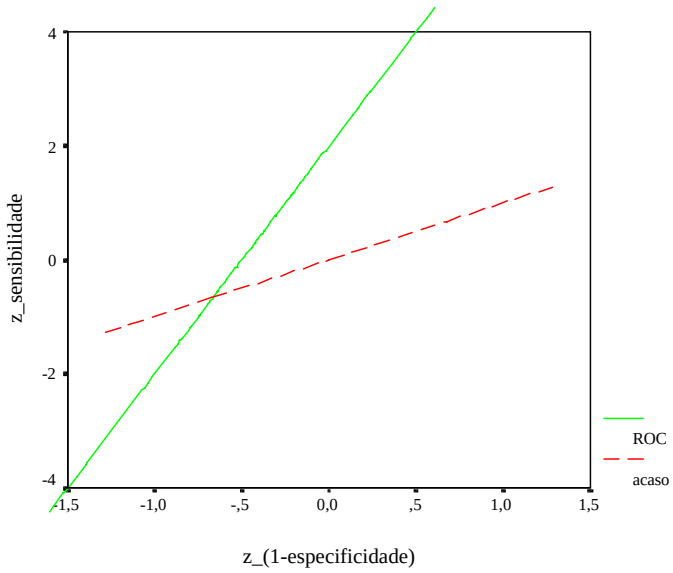


Figura 5.8: Representação da curva ROC para a situação descrita em b), no plano *binormal*.

Da análise do gráfico da figura 5.5, verifica-se que o cruzamento com a diagonal positiva ocorre para valores mais baixos do valor de corte, o que é indicativo da existência de uma certa proporção de indivíduos classificados como anormais para baixos valores de corte.

Na figura 5.6, é notória a relação linear devida à normalidade e o declive inferior a um devido à razão de desvios criada ($declive = \frac{\sigma_N}{\sigma_A} = 0.25$).

A situação ilustrada nos gráficos da figura 5.7 e 5.8, corresponde à razão $\frac{\sigma_N}{\sigma_A} = 4$. O cruzamento com a diagonal positiva verificado na figura 5.7 decorre de uma situação em que existe uma certa quantidade de indivíduos classificados como normais para elevados valores de corte.

Na figura 5.8, a relação linear devido à normalidade é evidente, e o declive superior a um também.

5.1.2 Função densidade de probabilidade Logística de igual variância

A função densidade de probabilidade logística é semelhante à função densidade de probabilidade normal (Gaussiana), e as funções densidade de probabilidade logísticas de igual variância para os casos denominados normais e para os casos denominados anormais conduzem a curvas ROC semelhantes às produzidas pelas funções densidade de probabilidade normais de igual variância [80].

A função densidade de probabilidade logística pode ser representada pela seguinte expressão com média μ, e variância $\frac{\pi^2\beta^2}{3}$ [63]:

$$f(x) = \frac{\exp\left[-\frac{(x-\mu)}{\beta}\right]}{\beta\left\{1+\exp\left[-\frac{(x-\mu)}{\beta}\right]\right\}^2} \qquad \begin{array}{l} -\infty < x < +\infty \\ -\infty < \mu < +\infty, \quad \beta > 0 \end{array} \tag{5.13}$$

As coordenadas da curva ROC determinadas para as hipóteses formuladas para as distribuições logísticas, e para um dado valor de corte c, serão dadas por:

$$\begin{aligned} FVP &= P(H_1|h_1) \\ &= \int_c^{+\infty} f(x|h_1)\, dx = 1 - \left[1 + \exp\left(-\frac{c-\mu_A}{\beta}\right)\right]^{-1} \end{aligned} \tag{5.14}$$

$$\begin{aligned} FFP &= P(H_1|h_0) \\ &= \int_c^{+\infty} f(x|h_0)\, dx = 1 - \left[1 + \exp\left(-\frac{c-\mu_N}{\beta}\right)\right]^{-1} \end{aligned} \tag{5.15}$$

Rearranjando as equações 5.14 e 5.15, concluí-se que a relação entre estas duas probabilidades para a distribuição logística é dada pela seguinte expressão:

$$FVP = \frac{FFP \exp\left(-\frac{\mu_N-\mu_A}{\beta}\right)}{1 - \left(1 - \exp\left(-\frac{\mu_N-\mu_A}{\beta}\right)\right) FFP} \tag{5.16}$$

que caracteriza a forma funcional da curva ROC no plano unitário, quando as funções densidade de probabilidade associadas aos dados são ambas logísticas com a mesma variância.

As figuras 5.9 e 5.10 ilustram a forma das curvas ROC quando as funções densidade de probabilidade são logísticas para valores fixos de $\mu_N = 50$ e $\beta_N = \beta_A = \beta = 5$, fazendo $\mu_A = 55$, $\mu_A = 60$ e $\mu_A = 70$, nos plano ROC unitário e no plano binormal, respectivamente.

Verifica-se assim que a forma das curvas ROC é semelhante ao caso em que se considera as duas funções densidade de probabilidade normais de igual variância. Na representação no plano *binormal* (figura 5.10) pode-se ver, contudo, que a representação ainda é aproximadamente uma recta, notando-se, no entanto, à medida que a distância entre as médias aumenta, uma concavidade, verificando-se para $d' = 4$ uma concavidade bastante acentuada.

5.1.3 Funções densidade de probabilidade Exponenciais negativas

Assumindo que a variável x tem distribuição Exponencial negativa, com parâmetro θ_N sob h_0 e com parâmetro θ_A sob $h_1$as expressões para as funções densidade de probabilidade serão:

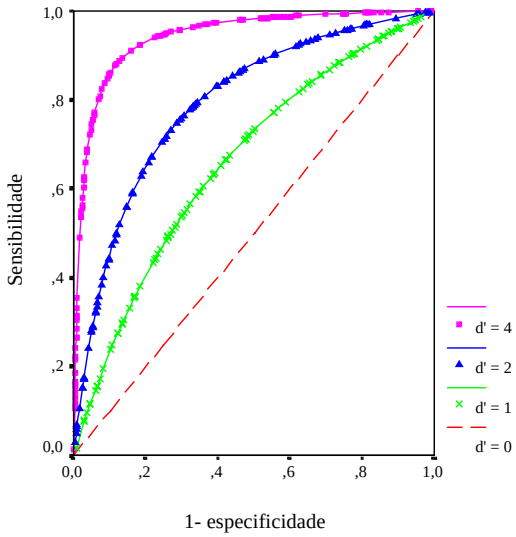


Figura 5.9: Representação das curvas ROC para distribuições Logísticas de igual variância no plano ROC.

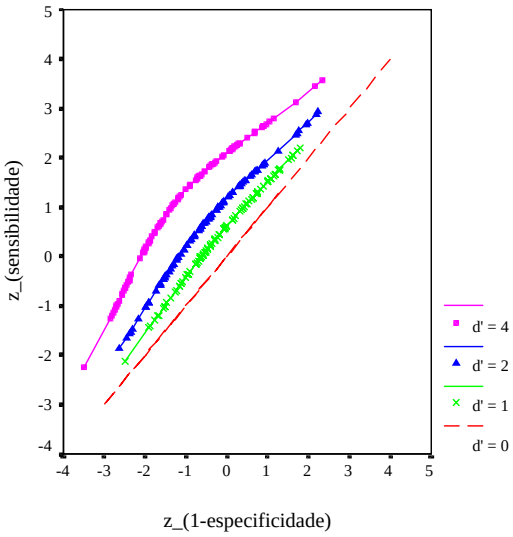


Figura 5.10: Representação das curvas ROC para distribuições Logísticas de igual variância no plano *binormal*.

$$f(x|h_0) = \frac{1}{\theta_N} \exp\left(-\frac{x}{\theta_N}\right) \quad \theta_N > 0, \quad x > 0 \tag{5.17}$$

$$f(x|h_1) = \frac{1}{\theta_A} \exp\left(-\frac{x}{\theta_A}\right) \quad \theta_A > 0, \quad x > 0 \tag{5.18}$$

As coordenadas da curva ROC determinadas para as hipóteses formuladas para as distribuições Exponenciais, e para um dado valor de corte c, serão dadas neste caso por:

$$FVP = P(H_1|h_1) = \int_c^{+\infty} f(x|h_1)\, dx = \exp\left(-\frac{c}{\theta_A}\right) \tag{5.19}$$

$$FFP = P(H_1|h_0) = \int_c^{+\infty} f(x|h_0)\, dx = \exp\left(-\frac{c}{\theta_N}\right) \tag{5.20}$$

Rearranjando as equações 5.19 e 5.20, concluí-se que a relação entre estas duas probabilidades para a distribuição Exponencial negativa é dada pela

seguinte expressão:

$$FVP = FFP^{\theta_N/\theta_A} \tag{5.21}$$

que caracteriza a forma funcional da curva ROC no plano unitário, quando as funções densidade de probabilidade associadas aos dados são ambas Exponenciais negativas, e com a condição $\theta_A > \theta_N$.

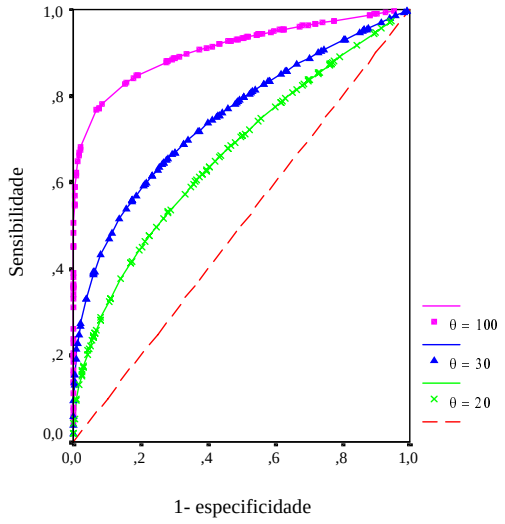


Figura 5.11: Representação das curvas ROC para distribuições Exponenciais negativas no plano ROC.

As figuras 5.11 e 5.12 ilustram a forma das curvas ROC quando as funções densidade de probabilidade são Exponenciais negativas para um valor fixo de $\theta_N = 10$, fazendo $\theta_A = 20, \theta_A = 30$ e $\theta_A = 100$, no plano ROC unitário e no plano *binormal*, respectivamente.

A partir da figura 5.11, verifica-se que a forma das curvas ROC pouco se altera em relação aos dois casos considerados anteriormente (normais de igual variância e logísticas de igual variância). No entanto, da análise da figura

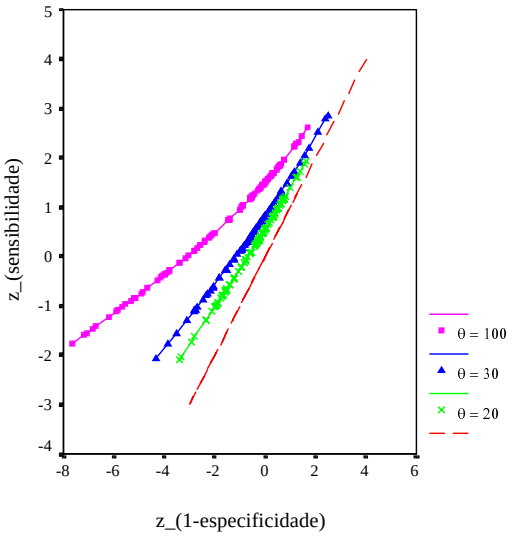


Figura 5.12: Representação das curvas ROC para distribuições Exponenciais negativas no plano *binormal*.

5.12, e como resultado da relação não linear entre FVP e FFP - expressão 5.21 - constata-se que a representação no plano *binormal* não é linear.

5.1.4 Funções densidade de probabilidade Uniformes num intervalo (a, b)

Assumindo que a variável x tem distribuição Uniforme, especificamente, sob h_0 , x tem distribuição Uniforme no intervalo (a_0, b_0) e sob a hipótese h_1, x tem distribuição Uniforme no intervalo (a_1, b_1), as formas para as funções densidade de probabilidade respectivas, serão:

$$f(x|h_0) = \frac{1}{b_0 - a_0} \qquad a_0 \leq x \leq b_0 \tag{5.22}$$

$$f(x|h_1) = \frac{1}{b_1 - a_1} \qquad a_1 \leq x \leq b_1 \tag{5.23}$$

As coordenadas da curva ROC determinadas para as hipóteses formuladas para as distribuições Uniformes, e para um dado valor de corte c, serão dadas neste caso por:

$$FVP = P(H_1|h_1) = \int_c^{b_1} f(x|h_1)\ dx = \frac{b_1 - c}{b_1 - a_1} \tag{5.24}$$

$$FFP = P(H_1|h_0) = \int_c^{b_0} f(x|h_0)\ dx = \frac{b_0 - c}{b_0 - a_0} \tag{5.25}$$

Rearranjando as equações 5.24 e 5.25, a relação entre estas duas probabilidades para a distribuição Uniforme é dada pela seguinte expressão:

$$FVP = \frac{b_1 - a_0}{b_1 - a_1} + \frac{b_0 - a_0}{b_1 - a_1}.FFP \tag{5.26}$$

que caracteriza a forma funcional da curva ROC no plano unitário, quando as funções densidade de probabilidade associadas aos dados são ambas Uniformes.

Nos estudos de simulação efectuados para a distribuição Uniforme, consideram-se as seguintes situações:

1. $X_N \sim U(0,4)$ e $X_A \sim U(2,6)$;
2. $X_N \sim U(0,2)$ e $X_A \sim U(0,4)$;
3. $X_N \sim U(0,1)$ e $X_A \sim U(0,4)$.

Os resultados obtidos encontram-se representados na figura 5.13 e 5.14 em termos de plano ROC e plano *binormal*, respectivamente.

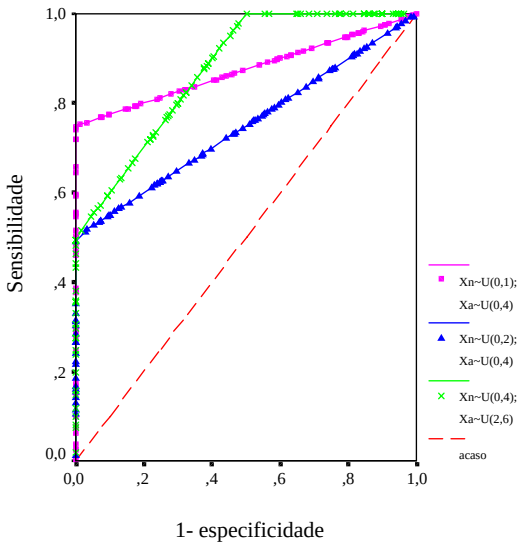


Figura 5.13: Representação das curvas ROC para distribuições Uniformes no plano ROC.

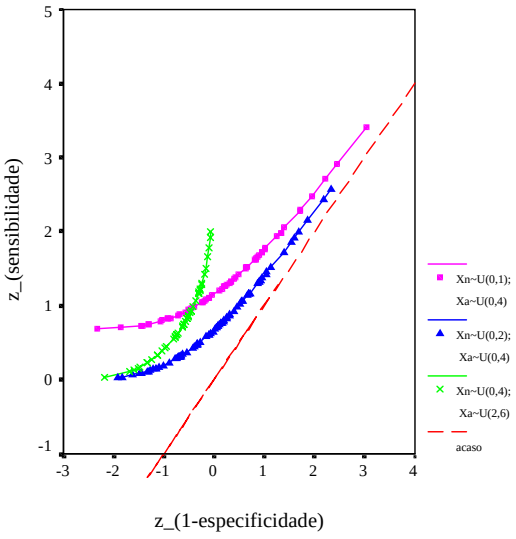


Figura 5.14: Representação das curvas ROC para distribuições Uniformes no plano *binormal*.

Da análise do gráfico da figura 5.13, verifica-se que a forma da curva ROC se altera por completo em relação aos casos descritos das normais, logísticas e exponenciais negativas. Quando se faz a representação no plano *binormal*, figura 5.14, torna-se evidente a relação não linear.

5.2 Cálculo do valor de área abaixo da curva ROC

Como mencionado por vários autores ([37], [58], [80]), a área abaixo da curva ROC é um dos índices mais utilizados para sumariar a "*qualidade*" da curva.

Como referido anteriormente, existem vários métodos para cálculo de áreas abaixo de uma curva ROC. Nesta secção, comparam-se os valores para o caso das função densidade de probabilidade Normais consideradas na secção 5.1, utilizando:

1. o declive e termo de intercepção da representação dos dados originais em papel de probabilidades *binormal* [58];
2. a aproximação à estatística U de Wilcoxon-Mann-Whitney [38].

5.2.1 Funções densidade de probabilidade Normais

Quando as funções densidade de probabilidades são Normais, existem duas situações a considerar, a da igualdade de variâncias e a da diferença de variâncias. Assim, para o caso de normais de igual variância irão corresponder declives unitários no plano *binormal*. Os resultados obtidos através do método de regressão linear simples, encontram-se na tabela 5.1:

Tabela 5.1: Comparação de áreas abaixo da curva ROC

μ_A	a	b	r^2	$\frac{a}{\sqrt{1+b^2}}$	$A_z = \Phi\left(\frac{a}{\sqrt{1+b^2}}\right)$	*A (WMW)*
55	1	1	1.0	0.707	0.760	0.744
60	2	1	1.0	1.414	0.921	0.916
70	4	1	1.0	2.828	0.998	0.9996

No caso das funções densidade de probabilidade para os casos designados por anormais e para os casos designados por normais terem variâncias diferentes ($\sigma_N \neq \sigma_A$), foram consideradas as seguintes situações:

a) $\frac{\sigma_A}{\sigma_N} = 4 > 1$;

b) $\frac{\sigma_A}{\sigma_N} = 0.25 < 1$.

Da mesma forma, utilizando a regressão linear e simples, obtiveram-se os resultados descritos na tabela 5.2

Tabela 5.2: Comparação de áreas abaixo da curva ROC para funções densidade de probabilidade Normais com variâncias diferentes.

Modelo	a	b	r^2	$\frac{a}{\sqrt{1+b^2}}$	$A_z = \Phi\left(\frac{a}{\sqrt{1+b^2}}\right)$	*A (WMW)*
a)	1.689	3.451	0.881	0.470	0.681	0.705
b)	0.475	0.254	0.921	0.460	0.677	0.604

Na construção dos modelos de regressão linear simples, o método utilizado para a estimação dos parâmetros a e b, foi o método dos mínimos quadrados.

Da análise dos resultados da tabela 5.1, verifica-se que os valores obtidos para A_z e A não variam muito, sendo apenas verificada alguma diferença a nível da segunda e terceira casa decimal. Assim, será lícito afirmar que o

cálculo do índice área abaixo da curva ROC, para o caso em que as duas funções densidade de probabilidade são Normais com igual variância, poderá ser feito quer através da aproximação à estatística de Wilcoxon-Man-Whitney, quer através da aproximação no plano *binormal* em que as rectas são bem definidas com declive unitário.

Na tabela 5.2, verifica-se os valores obtidos para A_z e A apresentam maior variação, sendo no caso a) ao nível da primeira casa decimal. Nesta situação, apesar da normalidade, que é indicada pela linearidade no plano *binormal*, existe diferença em termos dos desvios (declive não unitário), e o ajuste em termos de r^2 não é tão bom como nas situações descritas anteriormente na tabela 5.1.

5.2.2 Funções densidade de probabilidade Logística de igual variância

Para efectuar a análise de regressão linear simples, consideraram-se os modelos traçados no gráfico da figura 5.10. Os resultados obtidos encontram-se resumidos na tabela 5.3.

Tabela 5.3: Comparação de áreas abaixo da curva ROC para funções densidade de probabilidade Logísticas com a mesma variância.

μ_A	a	b	r^2	$\frac{a}{\sqrt{1+b^2}}$	$A_z = \Phi\left(\frac{a}{\sqrt{1+b^2}}\right)$	A *(WMW)*
55	0.560	0.983	0.996	0.399	0.655	0.734
60	1.078	0.970	0.983	0.774	0.781	0.687
70	2.019	0.953	0.953	1.461	0.928	0.936

Da análise dos resultados da tabela 5.3, verifica-se que a maior diferença nos valores de A_z e A regista-se nos dois primeiros casos. Apesar do valor

de r^2 indicar uma boa qualidade do ajuste, a análise do gráfico da figura 5.10, revela a não existência de linearidade, pelo que o método de estimação do índice área abaixo da curva ROC através de A_z, poderá não ser o mais indicado.

5.2.3 Funções densidade de probabilidade Exponenciais negativas

Para o caso das funções densidade Exponenciais negativas, na análise de regressão linear simples, consideraram-se os modelos traçados no gráfico da figura 5.12. Os resultados obtidos encontram-se resumidos na tabela 5.4.

Tabela 5.4: Comparação de áreas abaixo da curva ROC para funções densidade de probabilidade Exponenciais negativas.

θ_A	a	b	r^2	$\frac{a}{\sqrt{1+b^2}}$	$A_z = \Phi\left(\frac{a}{\sqrt{1+b^2}}\right)$	*A (WMW)*
20	0.558	0.800	0.999	0.436	0.669	0.643
30	0.850	0.716	0.996	0.691	0.755	0.767
100	1.507	0.458	0.992	1.370	0.915	0.946

Na tabela 5.4, verifica-se que os resultados obtidos para o índice área abaixo da curva ROC, A_z e A, apenas apresentam diferenças a nível da segunda casa decimal, e os valores de r^2 indicam uma boa qualidade do ajuste. No entanto, tendo em conta a informação fornecida pelo gráfico da figura 5.12, a relação linear não é tão evidente, pelo que o método de estimação preferível nesta situação seria através da aproximação à estatística de Wilcoxon-Mann-Whitney.

5.2.4 Funções densidade de probabilidade Uniformes num intervalo (a, b)

Para efectuar a análise de regressão linear simples, consideraram-se os modelos traçados no gráfico da figura 5.14, que correspondem às seguintes situações criadas:

1. $X_N \sim U(0,4)$ e $X_A \sim U(2,6)$
2. $X_N \sim U(0,2)$ e $X_A \sim U(0,4)$
3. $X_N \sim U(0,1)$ e $X_A \sim U(0,4)$.

Os resultados obtidos encontram-se resumidos na tabela 5.5.

Tabela 5.5: Comparação de áreas abaixo da curva ROC para funções densidade de probabilidade Uniformes num intervalo (a,b).

Modelo	a	b	r^2	$\frac{a}{\sqrt{1+b^2}}$	$A_z = \Phi\left(\frac{a}{\sqrt{1+b^2}}\right)$	*A (WMW)*
1	1.463	0.925	0.831	1.074	0.859	0.910
2	0,788	0.598	0.950	0.676	0.751	0.692
3	1.282	0.516	0.918	1.139	0.873	0.864

No caso dos valores obtidos de A_z e A para as densidades Uniformes consideradas, registados na tabela 5.5, verifica-se que existem diferenças a nível da primeira casa decimal (modelo 1 e 2). Se se analisar os valores de r^2, verifica-se que a qualidade do ajuste é pior em relação aos casos descritos nas secções anteriores.

De uma forma geral, pode-se concluir que se as densidades não forem normais, a estimação do índice área abaixo da curva ROC através de A_z deverá ser cautelosa e analisada em conjunto com a representação da curva ROC no plano *binormal.*

5.3 Relação entre o valor de área abaixo da curva ROC e a distribuição associada aos dados

O objectivo deste estudo, foi verificar qual a variação do índice área abaixo da curva ROC, com os parâmetros centrais e/ou dispersão das funções densidade de probabilidade associadas aos dados.

Para estudar a variação do índice área abaixo da curva ROC para diferentes valores dos parâmetros centrais e/ou dispersão consideraram-se as distribuições Exponencial negativa e Normal. A razão da escolha destas distribuições, é que a primeira apresenta caudas pesadas e a segunda é simétrica.

Nas secções que se seguem, são descritos alguns procedimentos experimentais tidos em conta, assim como os resultados obtidos.

5.3.1 Distribuições normais

Neste estudo considerou-se que a distribuição associada aos dados era Normal com média μ e variância σ^2, cuja densidade pode ser expressa por:

$$f(x) = \frac{1}{\sigma\sqrt{2\,\pi}} \exp\left[-\frac{(x-\mu)^2}{2\,\sigma^2}\right] \qquad \begin{array}{l} -\infty < x < +\infty \\ -\infty < \mu < +\infty, \quad \sigma > 0 \end{array}$$

Com base no pressuposto que a distribuição para os casos denominados normais era Normal com parâmetros (μ_N, σ_N^2), e para os casos denominados anormais era também Normal com parâmetros (μ_A, σ_A^2), realizaram-se as experiências descritas nas secções seguintes.

Igual parâmetro de dispersão,

Considerando $\sigma_N = \sigma_A = 5$, e amostras com dimensão $n_A = n_N = 50$, efectuaram-se três experiências, começando por fixar $\mu_N = 50$ e fazendo variar o valor de μ_A de 50, 60 e 70. Para cada conjunto gerado repetiu-se a experiência dez vezes.

Realizou-se o mesmo tipo de estudo para amostras de dimensão $n_A = n_N = 100$ e $n_A = n_N = 500$. Os resultados obtidos em termos de valores médios, para a área abaixo da curva ROC, A e erro padrão associado a esta, $SE(A)$ (calculados pelo método sugerido por Hanley e McNeil [37]), encontram-se resumidos nas tabelas 5.6, 5.7 e 5.8, respectivamente.

Tabela 5.6: Resultados para a Normal com $n_A = n_N = 50$

	$\mu_A = 50$	$\mu_A = 60$	$\mu_A = 70$
A	0.493	0.929	0.997
$SE(A)$	0.058	0.037	0.006

Tabela 5.7: Resultados para a Normal com $n_A = n_N = 100$

	$\mu_A = 50$	$\mu_A = 60$	$\mu_A = 70$
A	0.485	0.929	0.998
$SE(A)$	0.041	0.019	0.003

Da análise conjunta dos resultados destas três tabelas, pode-se verificar que à medida que a diferença entre as médias μ_N e μ_A aumenta, o valor da

Tabela 5.8: Resultados para a Normal com $n_A = n_N = 500$

	$\mu_A = 50$	$\mu_A = 60$	$\mu_A = 70$
A	0.530	0.922	0.998
$SE(A)$	0.018	0.009	0.001

área abaixo da curva ROC também aumenta e o erro padrão diminui. Este aumento deve-se ao facto de as função densidade de probabilidade associadas aos dados se encontrarem menos sobrepostas, sendo de prever que o valor de área abaixo da curva ROC tenda para um quando estas se encontrarem completamente separadas, situação que indica poder discriminante perfeito.

Diferentes parâmetros centrais e de dispersão

Neste caso procurou-se fazer uma variação do parâmetro de dispersão e ver qual a influência no valor da área abaixo da curva ROC.

Fixou-se a dimensão da amostra $n_A = n_N = 100$ para $\mu_N = 50$ e $\mu_A = 60$, com $\sigma_A/\sigma_N > 1$ (ver figura 5.4), consideraram-se as seguintes situações:

(i) $\dfrac{\sigma_N}{\sigma_A} = \dfrac{5}{10}$;
(ii) $\dfrac{\sigma_N}{\sigma_A} = \dfrac{5}{15}$;
(iii) $\dfrac{\sigma_N}{\sigma_A} = \dfrac{5}{20}$.

Os resultados obtidos em termos de valores médios das experiências efectuadas, encontram-se resumidos na tabela 5.9.

Tabela 5.9: Resultados para as situações descritas

	(i)	**(ii)**	**(iii)**
A	0.812	0.726	0.698
$SE(A)$	0.031	0.036	0.037

Da análise dos resultados da tabela 5.9, verifica-se que à medida que σ_A aumenta em relação a σ_N o valor de área abaixo da curva ROC diminui. Esta diminuição deve-se ao facto de a função densidade de probabilidade para os casos considerados anormais se tornar mais "achatada", o que significa maior dispersão de valores e consequentemente maior área de sobreposição com a função densidade de probabilidade dos casos considerados normais.

5.3.2 Distribuições Exponenciais negativas

Assumindo que a variável x tem distribuição Exponencial negativa, de acordo com as expressões das equações (5.17) e (5.18), começou-se por fixar o valor de $\theta_N = 10$ para os casos denominados normais e fez-se variar o parâmetro para os casos denominados anormais, θ_A, de forma que a relação fosse do tipo:

$$\theta_A = k\ \theta_N.$$

Graficamente ter-se-ia uma situação semelhante à ilustrada na figura 5.15.

Consideraram-se amostras de igual dimensão ($n_A = n_N = 50$, $n_A = n_N = 100$ e $n_A = n_N = 500$), e geraram-se assim, Exponenciais negativas fixando

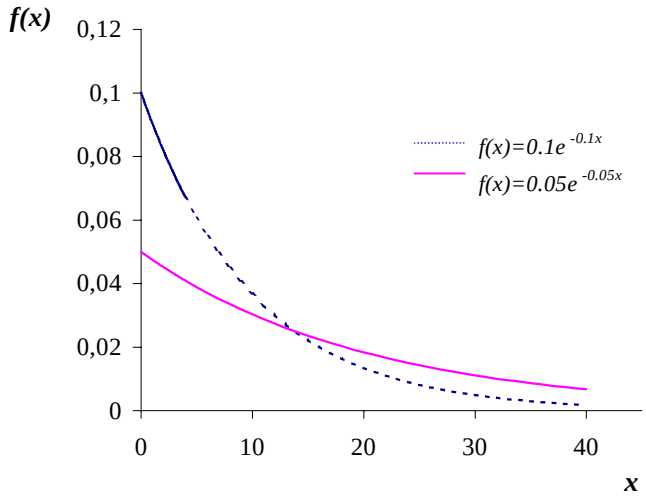


Figura 5.15: Sobreposição de duas funções densidade de probabilidade Exponenciais negativas com diferentes parâmetros θ.

$\theta_N = 10$ e fazendo variar de $10, 15, 20$ e 100, repetindo-se o processo aleatório dez vezes.

Os resultados em termos de valores médios, obtidos neste estudo para a área abaixo da curva ROC, A e erro padrão associado a esta, $SE(A)$ (calculados pelo método sugerido por Hanley e McNeil [37]), encontram-se resumidos nas tabela 5.10, 5.11 e 5.12.

Tabela 5.10: Resultados para a Exponencial negativa com $n_A = n_N = 50$

	$\theta_A = 10$	$\theta_A = 15$	$\theta_A = 20$	$\theta_A = 100$
A	0.543	0.607	0.705	0.907
$SE(A)$	0.058	0.056	0.052	0.031

Tabela 5.11: Resultados para a Exponencial negativa com $n_A = n_N = 100$

	$\theta_A = 10$	$\theta_A = 15$	$\theta_A = 20$	$\theta_A = 100$
A	0.489	0.622	0.664	0.909
$SE(A)$	0.041	0.039	0.038	0.022

Tabela 5.12: Resultados para a Exponencial negativa com $n_A = n_N = 500$

	$\theta_A = 10$	$\theta_A = 15$	$\theta_A = 20$	$\theta_A = 100$
A	0.498	0.609	0.669	0.912
$SE(A)$	0.018	0.018	0.017	0.010

Da análise conjunta dos resultados nestas três tabelas, verifica-se que de uma forma geral à medida que o parâmetro θ_A aumenta, o valor da área abaixo da curva ROC também aumenta, o que significa maior poder discriminante e revela o maior afastamento das duas curvas, consequentemente menor área de sobreposição entre as duas densidades. E para $\theta_N = \theta_A$, o valor da área abaixo da curva ROC está próximo de 0.5, o que significa a não existência de poder discriminante, isto é, as curvas encontram-se sobrepostas.

Verifica-se também, que quanto maior for o valor de k, mais significativo é o aumento no valor de área abaixo da curva ROC, o que pode ser explicado pela forma da função densidade de probabilidade da Exponencial negativa, como ilustrado na figura 5.15.

Por outro lado, o aumento da dimensão das amostras faz baixar significativamente os valores dos erros padrão, que passam a ser da ordem de 1% para

amostras de grande dimensão ($n_A = n_N = 500$), não se verificando contudo grandes diferenças em termos de valores de área abaixo da curva ROC nas três tabelas para todos os casos.

5.4 Discussão

A generalidade dos estudos sobre curvas ROC assume a hipótese da Normalidade para a distribuição dos dados. Para curvas ROC empíricas, geradas a partir de dados amostrais, autores como Swets [80] e Metz [57] referem que, no plano *binormal*, estas podem ser aproximadas por rectas. Na secção 5.1, várias curvas ROC foram geradas a partir de distribuições não normais. Como se pode ver pelas figuras apresentadas, com excepção do caso Uniforme, as curvas ROC no plano unitário não parecem apresentar grandes diferenças de forma. No entanto, a representação no plano *binormal* mostra claramente uma relação não linear. Apesar de certos segmentos destas curvas poderem apresentar um comportamento linear, a aproximação por uma recta no plano *binormal*, só é justificada pela prática generalizada e por amostras de pequena dimensão. Ora, tal aproximação pode induzir conclusões erradas quanto à forma da distribuição subjacente aos dados.

Na secção 5.2, a análise das tabelas 5.2 - 5.5, revela duma forma geral, bons ajustes de rectas no plano *binormal* em termos de r^2 ($r^2 > 0.8$), notando-se, no entanto, que os valores mais baixos do coeficiente de determinação se verificam para um dos modelos normais com variâncias diferentes ($\sigma_A > \sigma_N$) e para um modelo da Uniforme ($X_N \sim U(0,4)$ e $X_A \sim U(2,6)$). Contudo, uma análise pormenorizada de resíduos permitiria averiguar que o padrão destes não é aleatório, nomeadamente no caso da Uniforme, pelo que a "*qualidade*" do ajuste linear deverá ser analisado conjuntamente, em

termos de r^2 e padrão de resíduos.

Este estudo averiguou, também, a diferença existente no valor do índice área abaixo da curva ROC, utilizando dois tipos de abordagem, a abordagem paramétrica que considera o ajuste dos dados ROC num plano *binormal*, e a abordagem não paramétrica que considera a aproximação deste índice ao valor da estatística de Wilcoxon-Mann-Whitney.

Apesar dos valores determinados de A_z e A não variarem muito, verificando-se em alguns casos apenas diferenças a nível da segunda casa decimal, neste estudo optar-se-à pela abordagem não paramétrica, pois garante a inexistência de pressupostos distribucionais associados aos dados. Por outro lado, em termos matemáticos, a estatística de Wilcoxon-Mann-Whitney envolve menos cálculos, tornando-se num processo mais simples. A abordagem paramétrica, envolve um processo de cálculo mais complexo, para além de se ter de estimar os parâmetros associados à recta de ajuste (o termo de intercepção e o declive da recta) e verificar a qualidade do ajuste linear.

Na secção 5.3 verifica-se, dum modo geral, que o aumento de área abaixo da curva ROC é mais significativo quando se faz variar os parâmetros de localização da Normal do que na variação do parâmetro da Exponencial negativa. Este aumento deve-se à forma específica das função densidade de probabilidade da Normal e da exponencial negativa, que no caso da Normal permite maior sobreposição das funções densidade de probabilidade quando as médias se aproximam, e menor quando estas se afastam.

Este estudo permitiu analisar a diferença de comportamento, em termos de índice área abaixo da curva ROC, numa situação em que as funções densidade de probabilidade são simétricas, de uma outra em que as funções densidade de probabilidade possuem caudas pesadas.

Capítulo 6

Aplicações

Como descrito em capítulos anteriores, o índice área abaixo da curva ROC é uma medida sumária da curva ROC, que pode ser utilizada para avaliar o desempenho de um sistema de diagnóstico. A principal vantagem deste índice é que a sua utilização não depende da distribuição associada aos dados.

A aplicabilidade da análise ROC, através do estudo das curvas ROC é muito vasta. Salientam-se áreas como a psicologia, o controle de qualidade, a medicina, a imagem radiológica, entre outras.

Neste capítulo procura-se ilustrar a utilização desta técnica através de alguns exemplos, nomeadamente no campo da medicina, para o qual se dispôs de dados gentilmente cedidos, pelo serviço de Neonatologia do Hospital Garcia de Orta de Almada.

6.1 A avaliação do risco de morte em recém-nascidos de muito baixo peso - amostras relacionadas

O peso do recém-nascido foi durante muito tempo a medida mais importante de risco neonatal inicial, sobretudo devido à sua importância e facilidade de avaliação. Contudo, começaram a ser necessárias formas de avaliação mais precisas para o risco de mortalidade neonatal inicial, permitindo assim a comparação entre serviços, regiões e mesmo países.

Classicamente, as taxas de mortalidade neonatal são consideradas um dos indicadores mais importantes para a avaliação do desempenho dos cuidados de saúde e do estágio de desenvolvimento da própria sociedade. Cada vez mais, os recém-nascidos de muito baixo peso (menos de 1500 gramas ao nascer) contribuem de forma significativa para as taxas de mortalidade e morbilidade.

Nos últimos anos foram desenvolvidas escalas de gravidade clínica com este objectivo. Dessas escalas, salientam-se o *CRIB (Clinical Risk Index for Babies), NTISS (Neonatal Therapeutical Intervention Score System), SNAP (Score for Neonatal Acute Physiology)* e *SNAP-PE (Score for Neonatal Acute Physiology - Perinatal Extension*). De notar que estes diferentes sistemas de pontuação implicam a recolha de variáveis ao longo de determinado período de tempo. Assim, e para os sistemas referidos, o número de variáveis a recolher varia entre 6 (*CRIB*), 26 (*SNAP*), 29 (*SNAP-PE*) e 48 (*NTISS*).

Todas estas variáveis são recolhidas nas primeiras 24 horas de vida, sendo, excepcionalmente para o *CRIB*, o período reduzido para as 12 horas posteriores ao parto. Por esta razão, o *CRIB* torna-se num índice mais fácil de ser usado, quer em termos de tempo, quer em termos do número de variáveis.

O diagnóstico como processo imperfeito que é, conduz a que num teste em que se pretende classificar os indivíduos em anormais e normais exista sempre a possibilidade de cometer um de dois tipos de erros: classificar um indivíduo anormal como normal e, vice-versa, classificar um indivíduo normal como anormal.

Este estudo teve como principal objectivo, comparar, no mesmo conjunto de indivíduos, quatro tipos diferentes de índices de gravidade clínica para determinação do risco de morte para recém-nascidos de muito baixo peso. Os quatro índices estudados foram, *CRIB, NTISS, SNAP* e *SNAP-PE*. Inclui-se ainda, a variável *PESO* por esta ser também uma medida de risco neonatal.

6.1.1 Testes de hipóteses

Sob o ponto de vista clínico existe a necessidade de avaliar duma forma precisa o risco de mortalidade neonatal inicial para os recém-nascidos de muito baixo peso, pois este grupo contribui de forma significativa para as taxas de mortalidade e morbilidade. Assim, uma escala de gravidade clínica com elevado poder discriminante entre dois estados (falecido e sobrevivo), poderá funcionar como índice indicativo do desempenho das unidades de cuidados intensivos neonatais, porque tem em conta outras diferenças no risco, nomeadamente aquelas que dizem respeito à severidade inicial da doença.

Neste teste de diagnóstico, as hipótese para o problema são:

H_0 : O recém-nascido vai falecer, **M**

H_1 : O recém-nascido vai sobreviver, **V**

Assim, de acordo com as expressões definidas para os erros de tipo I (equação (2.7)) e de tipo II (equação (2.8)), verifica-se que para um valor de corte específico a representação ROC dá a probabilidade de aceitar H_0, isto

é, considerar que o recém-nascido de muito baixo peso vai falecer.

6.1.2 Descrição dos Dados

A amostra em estudo é constituída por 169 recém-nascidos de muito baixo peso (menos de 1500 g) internados na Unidade de Neonatologia do Hospital Garcia de Orta. Esta recolha foi feita de um modo retrospectivo sobre a mesma amostra, por forma a permitir a comparação entre as diversas escalas, durante o período de três anos, de 1992 a 1995. Dos 169 recém-nascidos de muito baixo peso em estudo, 133 sobreviveram , tendo-se observado 36 óbitos.

Como foi referido, para além das escalas de gravidade em estudo, o peso do recém-nascido foi também incluído como uma escala de gravidade per si, através do agrupamento em nove classes. As classes foram determinadas tendo em consideração a gama de valores observados (entre 540 g e 1500 g); o seu número foi calculado por forma a garantir a maior área abaixo da curva ROC em relação à escala contínua.

6.1.3 Resultados

Nas figuras 6.1, 6.2, 6.3, 6.4 e 6.5 estão representados os gráficos de distribuição de frequências para os recém-nascidos de baixo peso que faleceram e para os que sobreviveram.

Como se pode verificar, em todos estes gráficos existe uma sobreposição das distribuições para os recém-nascidos de baixo peso falecidos e sobrevivos. A análise gráfica mostra também que as diversas escalas apresentam diferentes graus de sobreposição das distribuições de falecidos e sobrevivos.

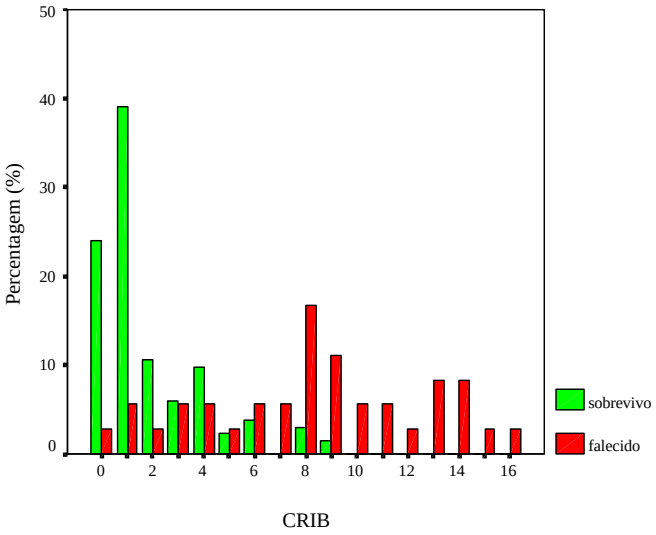


Figura 6.1: Distribuição de frequências para os recém-nascidos de baixo peso que faleceram e para os que sobreviveram em relação ao *CRIB*.

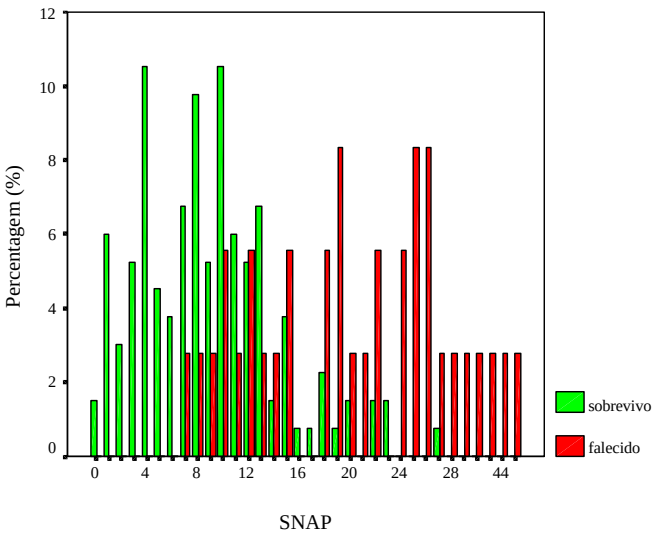


Figura 6.2: Distribuição de frequências para os recém-nascidos de baixo peso que faleceram e para os que sobreviveram em relação ao *SNAP*.

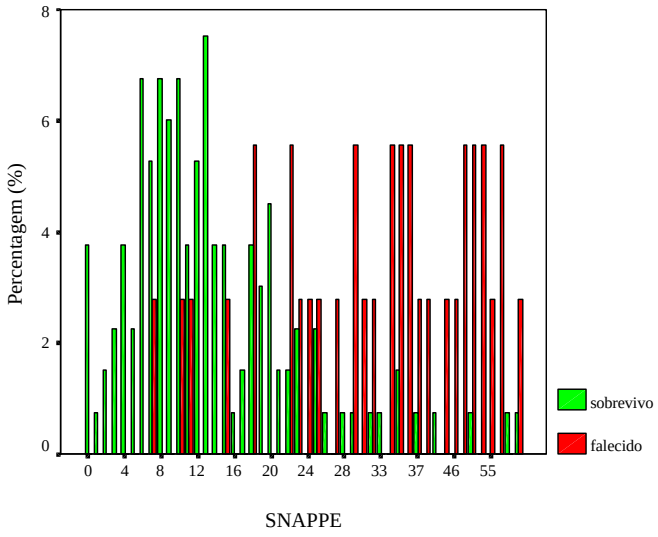


Figura 6.3: Distribuição de frequências para os recém-nascidos de baixo peso que faleceram e para os que sobreviveram em relação ao *SNAP-PE*.

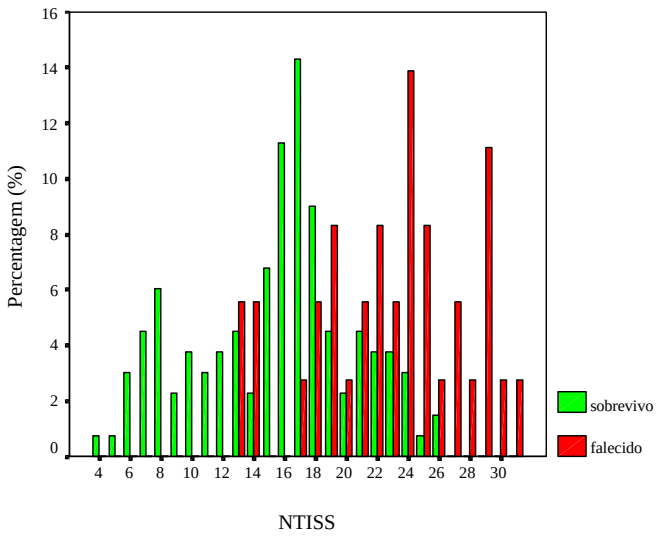


Figura 6.4: Distribuição de frequências para os recém-nascidos de baixo peso que faleceram e para os que sobreviveram em relação ao *NTISS*.

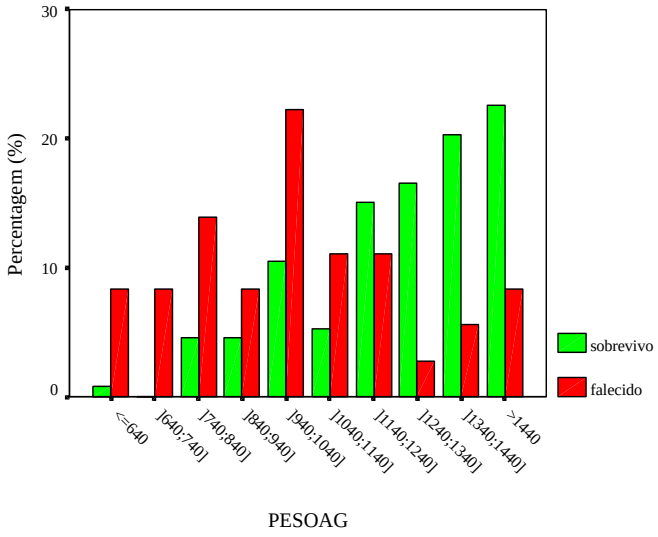


Figura 6.5: Distribuição de frequências para os recém-nascidos de baixo peso que faleceram e para os que sobreviveram em relação ao *PESOAG*.

Desta análise verifica-se ainda que, para os índices *CRIB, SNAP, SNAP-PE* e *NTISS*, valores elevados da variável de decisão, x, tendem a indicar a ocorrência de falecimento. Com a variável *PESOAG* verifica-se precisamente o contrário, dado que se verifica uma maior taxa de sobrevivência entre os bebés de peso mais elevado.

Curvas ROC

A metodologia utilizada para o cálculo das áreas abaixo das curvas ROC (A), e respectivos erros padrão ($SE(A)$), foi a aproximação não paramétrica à estatística de Wilcoxon-Mann-Whitney (equação (4.31)) sugerida por Hanley e McNeil [37]. Na tabela 6.1, encontra-se o resumo destes valores para os vários índices.

Tabela 6.1: Valores de A e $SE(A)$ para os diferentes índices na previsão de falecimento para os recém-nascidos de muito baixo peso.

Índice	**Área abaixo da curva ROC** (A)	**Erro padrão** ($SE(A)$)
CRIB	0.90	0.03
PESOAG	0.77	0.05
SNAP	0.88	0.03
SNAPPE	0.88	0.03
NTISS	0.84	0.04

Como referido em capítulos anteriores, graficamente a curva ROC representa a probabilidade de um verdadeiro positivo em função da probabilidade de um falso positivo para uma gama de valores de corte. Neste estudo, traçaram-se as curvas ROC empíricas para os cinco índices no plano ROC unitário, como ilustrado na figura 6.6.

À curva ROC que se aproxima mais do canto superior esquerdo, corresponderá o índice que deve ser preferido para previsão do risco de morte para os recém-nascidos de baixo peso. No entanto, apesar do *CRIB* apresentar a maior área, existe uma dificuldade resultante do cruzamento das curvas.

Coeficientes de correlação

As matrizes de correlação para os recém-nascidos falecidos (r_A) e para os recém-nascidos sobrevivos (r_N) foram determinadas pelo tau Kendall e apresentam-se na tabela 6.2.

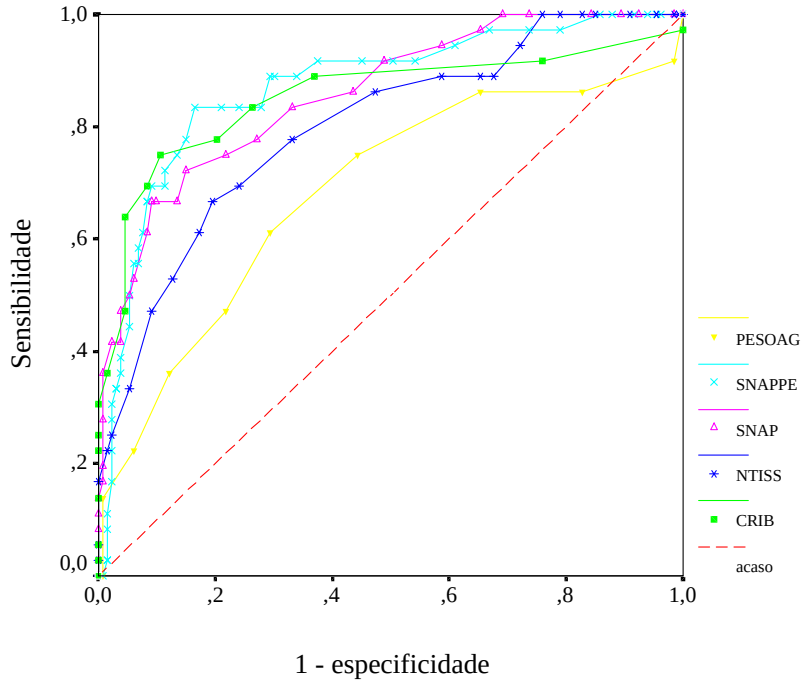


Figura 6.6: Gráfico das curvas ROC para os 5 índices.

Os valores do coeficiente de correlação, r_{HM} e r_{DL}, determinados pela metodologia de Hanley e McNeil [55] e a partir da rotina descrita por DeLong e DeLong em [22], encontram-se na tabela 6.3.

Tabela 6.2: Matrizes de correlação para os recém-nascidos falecidos (r_A) , e para os recém-nascidos sobrevivos (r_N).

CRIB	NTISS	SNAP	SNAPPE	$\leftarrow r_A \quad r_N \rightarrow$	CRIB	NTISS	SNAP	SNAPPE
-0.100	0.043	-0.053	-0.377	PESOAG	-0.368	-0.258	-0.203	-0.406
	0.169	0.365	0.263	CRIB		0.488	0.465	0.380
		0.250	0.143	NTISS			0.494	0.375
			0.585	SNAP				0.579

Tabela 6.3: Matrizes de correlação determinadas pela metodologia de DeLong e de Hanley e McNeil.

CRIB	NTISS	SNAP	SNAPPE	$\underleftarrow{r_{DL}} \quad \underrightarrow{r_{HM}}$	CRIB	NTISS	SNAP	SNAPPE
-0.011	0.101	0.126	0.545	PESOAG	-0.19	-0.09	-0.11	-0.35
	0.346	0.203	0.241	CRIB		0.27	0.34	0.25
		0.194	0.206	NTISS			0.31	0.205
			0.669	SNAP				0.495

Testes de comparação múltipla

Os resultados dos testes de comparação múltipla considerando os valores de r_{HM} e r_{DL} da tabela 6.3, encontram-se resumidos na tabela 6.4 em termos de *valor de prova*.

6.1.4 Discussão dos resultados

A comparação das áreas abaixo das curvas ROC, permite concluir que o melhor índice para avaliação do risco de falecimento em recém-nascidos de muito baixo peso é o *CRIB*, pois é aquele que apresenta maior valor de área ($A = 0.90$) com menor erro padrão ($SE(A) = 0.03$).

Tabela 6.4: Valores de prova para os testes de comparação múltipla entre os diferentes índices, pela metodologia de DeLong e de Hanley e McNeil.

CRIB	NTISS	SNAP	SNAPPE	$\underleftarrow{p_{DL}}$ $\underrightarrow{p_{HM}}$	CRIB	NTISS	SNAP	SNAPPE
0.027	0.183	0.045	0.005	PESOAG	0.038	0.226	0.056	0.084
	0.195	0.668	0.684	CRIB		0.219	0.637	0.682
		0.416	0.409	NTISS			0.380	0.409
			0.971	SNAP				0.977

No entanto, é de salientar que devido aos cruzamentos existentes entre as curvas, poder-se-ia ter efectuado o teste à fracção de verdadeiros positivos (sensibilidade) FVP, para um ponto particular de fracção de falsos positivos (1-especificidade) FFP_0, como descrito por Metz [58] (ver 4.10.1). Tal teste poderia ser justificado se o objectivo do estudo fosse avaliar o desempenho dos diferentes índices num ponto particular da fracção de falsos positivos. Através do gráfico da figura 6.6 verifica-se que, para valores baixos da escala do *CRIB*, poderá haver melhor desempenho dos outros índices (*SNAP, SNAPPE* e *NTISS*). Em termos clínicos, esta diferença não se justifica, dado que para valores baixos nas escalas (indicador de sobrevivência), todos deverão apresentar um bom desempenho. Este cruzamento do *CRIB* é explicado pela existência de uma determinada proporção de recém-nascidos de baixo peso que morre com um valor baixo de *CRIB* (gráfico da figura 6.1).

Verificou-se ainda, de uma forma geral, que a correlação existente entre os vários índices é mais significativa para os recém-nascidos sobrevivos do que para os recém-nascidos falecidos; tal poderá dever-se ao facto da dimensão da amostra de casos sobrevivos (133) ser significativamente mais elevada do que a de casos falecidos (36).

Da análise dos coeficientes de correlação, r, obtidos pelas duas metodologias, verifica-se que não existem diferenças em termos dos testes de comparação múltipla sendo no entanto a metodologia sugerida por DeLong e DeLong mais exacta, dado que utiliza a teoria das estatísticas *U-generalizadas* para estimação da matriz de covariâncias em vez de valores aproximados.

Resultante das comparações múltiplas, usou-se um nível de significância, $\alpha = 0.005$ o que garante globalmente um nível de 5%; a partir da tabela 6.4, não se verificaram diferenças significativas entre os índices.

É de salientar que o desenho retrospectivo deste estudo pode ter afectado o desempenho dos índices *SNAP, SNAPPE* e *NTISS*. Como nem todos os testes incluídos nestes índices são feitos rotineiramente, a sua inclusão por via do estudo pode não ser justificada.

Importa notar que, devido à natureza retrospectiva do estudo e ainda pelo facto de os dados relativos aos diferentes índices terem sido recolhidos sobre a mesma amostra, não é possível garantir a sua independência.

A avaliação do desempenho dos cinco índices estudados não foi conclusiva quanto ao que poderá apresentar melhor performance, no entanto a avaliar pela complexidade das escalas (em termos do número de variáveis a recolher e tempo de recolha) em relação ao *CRIB* e, pelo facto deste índice apresentar um maior valor de área abaixo da curva ROC e menor erro padrão, sugere-se que o *CRIB* poderá ser considerado o melhor índice indicativo do risco de mortalidade neonatal.

6.2 A Idade Gestacional como medida de prognóstico: análise através das curvas ROC para amostras relacionadas

A idade gestacional é considerada um factor de prognóstico muito importante de uma gravidez, quer no que diz respeito à mortalidade, quer relativamente ao aparecimento de doenças ou sequelas no bebé [6].

Em geral, os recém-nascidos com melhor prognóstico de sobrevivência possuem idades gestacionais elevadas. Por isso, a idade gestacional torna-se um factor importante na decisão de desencadear ou não um parto.

O método mais utilizado, e mais fiável para medição da idade gestacional é a ecografia. A *Idade Gestacional Obstétrica (IGO)* é a idade gestacional atribuída pela ecografia, corrigida por alguns factores associados com outros métodos de datação, como por exemplo, a data da última menstruação [65], [24], [5].

A *Idade Gestacional Neonatal (IGN)* avalia a idade gestacional após o nascimento, pois nem sempre se sabia a data da última menstruação, ou então, a idade gestacional atribuída não era compatível com o aspecto do bebé. Duma forma geral a *IGN* toma um valor superior em relação à *IGO*.

A análise de diagnóstico pretende, neste caso, determinar a influência da idade gestacional como factor de prognóstico no parto (bebé falecido ou sobrevivo), mas também comparar as duas medidas, *IGO* e *IGN,* avaliando se alguma das escalas é superior.

As hipóteses estatísticas formuladas são as mesmas da secção 6.1.1, e a metodologia a utilizar é a descrita na secção 4.10.2, dado que se tratam de amostras relacionadas.

Efectuou-se também, uma comparação através do *teste-t* para dados em-

parelhados, para tentar comprovar que na realidade a *IGN* toma valores superiores em relação à *IGO*

6.2.1 Descrição dos dados

A amostra em estudo é constituída por 223 bebés nascidos no Hospital Garcia de Orta, em Portugal. Esta recolha foi feita de um modo retrospectivo sobre a mesma amostra, por forma a permitir a comparação entre as duas idades gestacionais, durante o ano de 1995. Dos 223 bebés, 194 sobreviveram (classificados como sobrevivos), tendo sido registado 29 óbitos (classificados como falecidos). Foram ainda considerados os bebés de "risco", isto é, bebés com muito baixo peso à nascença (inferior a 1500 g), num total de 157, tendo sido observado 26 óbitos e 131 sobrevivos.

6.2.2 Resultados

Nas figuras 6.7 e 6.8 estão representadas as distribuições de frequências para os bebés que faleceram e para os que sobreviveram, em função da *IGO* e da *IGN*, considerando todos os bebés, e nas figuras 6.9 e 6.10 considerando apenas os recém-nascidos de muito baixo peso (inferior a 1500 g), respectivamente.

Como se pode verificar pela análise dos gráficos das figuras 6.7, 6.8, 6.9 e 6.10, as distribuições dos bebés sobrevivos e dos bebés falecidos sobrepõem-se. Pode ainda verificar-se que, de uma forma geral, valores elevados de idade gestacional tendem a indicar que o bebé vai sobreviver e valores baixos de idade gestacional tendem a indicar que o bebé irá falecer. Note-se ainda a existência de sobrevivos para baixos valores de idade gestacional.

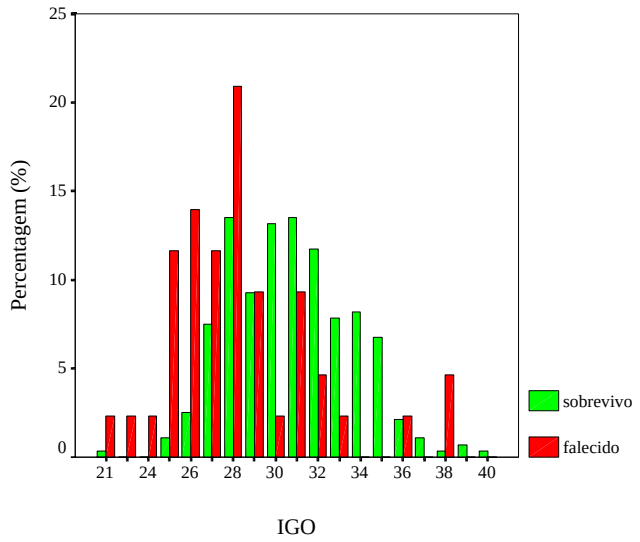


Figura 6.7: Distribuição de frequências para os bebés que faleceram e para os que sobreviveram em função da IGO, considerando todos os bebés.

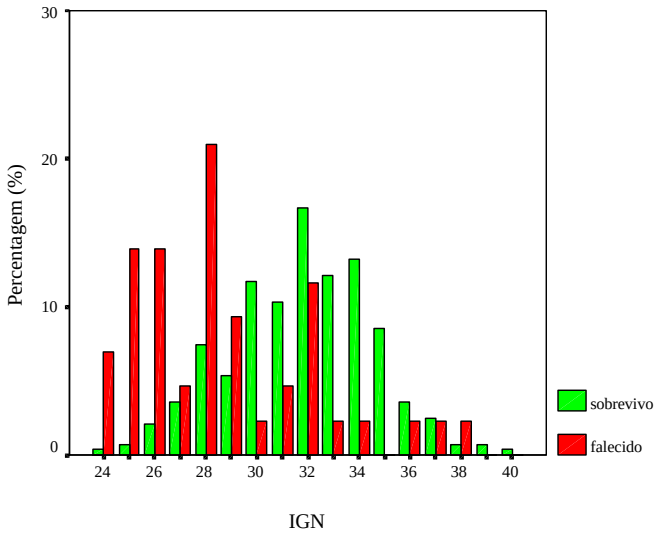


Figura 6.8: Distribuição de frequências para os bebés que faleceram e para os que sobreviveram em função da IGN, considerando todos os bebés.

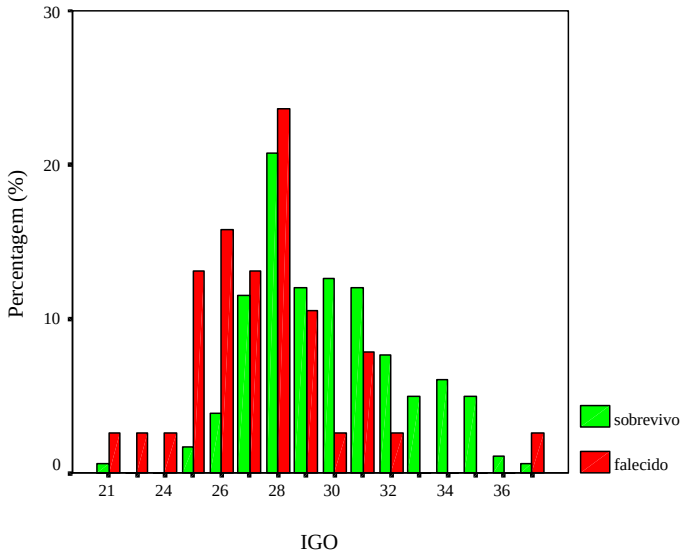


Figura 6.9: Distribuição de frequências para os bebés que faleceram e para os que sobreviveram em função da IGO (peso < 1500 g).

Áreas abaixo da curva ROC e erros padrão

Os valores observados da área abaixo da curva ROC e os respectivos erros padrão, para a *IGO* e a *IGN*, para o conjunto de todos os bebés e para aqueles que possuem peso abaixo de 1500 g, encontram-se resumidos na tabela 6.5.

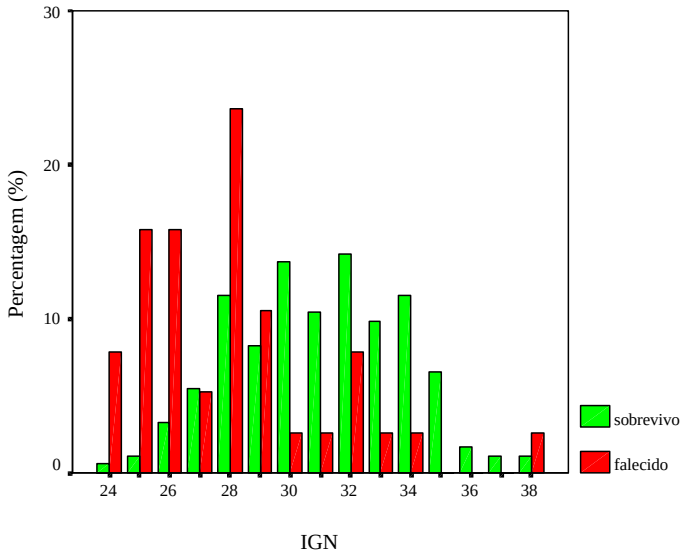


Figura 6.10: Distribuição de frequências para os bebés que faleceram e para os que sobreviveram em função da IGN (peso $<$ 1500 g).

Tabela 6.5: Valores de A e $SE(A)$ para todos os bebés e para aqueles com peso abaixo de 1500 g.

	Todos os bebés		**Bebés de peso $<$ 1500 g**	
	A	$SE(A)$	A	$SE(A)$
IGO	0.803	0.05	0.815	0.05
IGN	0.812	0.05	0.833	0.05

Curvas ROC

Na figura 6.11 e 6.12, representam-se as curvas ROC para a IGO e a IGN para todos os bebés e considerando apenas os recém-nascidos de muito baixo

peso (inferior a 1500 g), respectivamente.

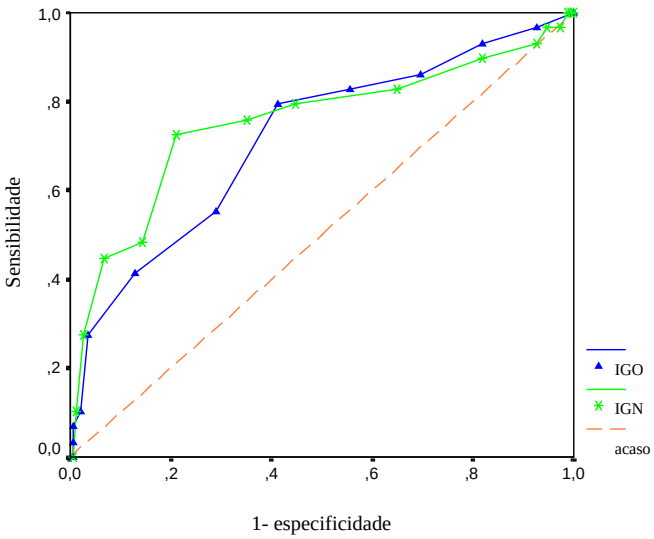


Figura 6.11: Curvas ROC para a IGO e para a IGN considerando todos os bebés.

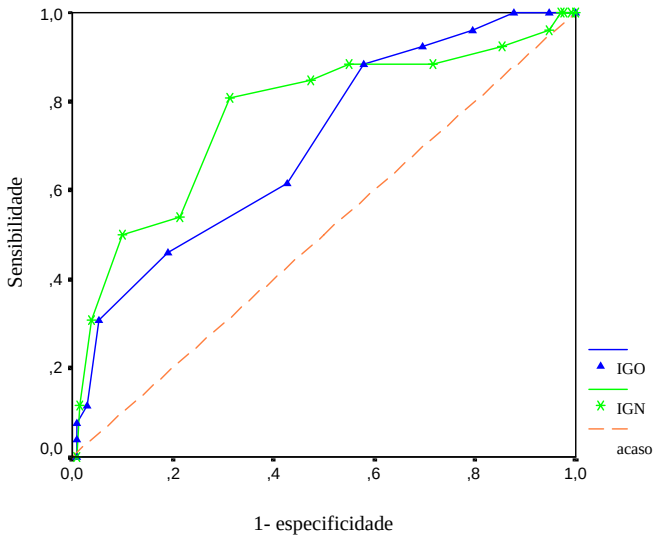


Figura 6.12: Curvas ROC para a IGO e para a IGN considerando os bebés com peso inferior a 1500 g.

Testes à diferença das áreas

(i) Matrizes de correlação Os valores do coeficiente de correlação para os bebés falecidos (r_A) foram de 0.874 e 0.877, considerando todos os bebés e somente os bebés de peso inferior a 1500 g, respectivamente. Para bebés sobrevivos (r_N) foram 0.544 e 0.507, para os dois estudos considerados. Estes coeficientes foram determinados pelo tau de Kendall. Os valores dos coeficientes de correlação obtidos da tabela de Hanley e McNeil (r_{HM}) foram 0.66 e 0.63, respectivamente. Calcularam-se ainda os valores dos coeficientes de correlação utilizando a metodologia de DeLong, obtendo-se $r_{DL} = 0.734$ e $r_{DL} = 0.649$, considerando todos os bebés e somente os bebés de peso inferior a 1500 g, respectivamente.

(ii) Testes de hipóteses Quer para o estudo considerando todos os bebés quer para o estudo em que se consideram apenas bebés de muito baixo peso, não se verificaram diferenças significativas entre a IGO e a IGN, em termos do índice área abaixo da curva ROC, tendo sido observados os *valores de prova* de 0.827 e 0.676, respectivamente, quando se utiliza a metodologia de Hanley e McNeil. Utilizando a metodologia de DeLong, os *valores de prova* observados foram 0.805 e 0.667, respectivamente.

6.2.3 Discussão dos resultados

A comparação das áreas abaixo das curvas ROC, para o estudo que considera todos os bebés e para o estudo que considera os bebés de peso inferior a 1500 g, nada permite concluir quanto à melhor medida de avaliação do risco de morte para os bebés nos dois estudos efectuados, tal como é confirmado pelos testes à diferença das áreas. No entanto, pelos valores de área abaixo da curva ROC, verifica-se que a idade gestacional pode ser considerada como um factor de prognóstico importante para a sobrevivência dos bebés, sendo assim um indicador importante na tomada de decisão sobre a indução de um parto.

A comparação entre os dois métodos de datação, *IGO* e *IGN*, permite concluir que a diferença média (1.26 semanas) é estatisticamente significativa ($t = 9.035$, $p < 0.01$).

6.3 Comparação de unidades de cuidados intensivos neonatais - amostras independentes.

A comparação do desempenho de unidades de cuidados intensivos neonatais baseada na sua taxa de mortalidade necessita de métodos exactos para ajustar as diferenças existentes no risco inicial dos seus pacientes, pois esta comparação terá de ter em conta não só os aspectos associados aos bebés nascidos na unidade, mas também quais as condições em que são recebidos os recém-nascidos provenientes de outras unidades. Assim, uma unidade de cuidados intensivos neonatais poderá apresentar uma elevada taxa de mortalidade e no entanto esta ser devida à recepção de recém-nascidos externos a esta unidade, com um risco inicial muito elevado.

Como o peso à nascença foi desde sempre uma medida importante na determinação do risco neonatal inicial, não houve necessidade de desenvolver novos sistemas de classificação para os cuidados intensivos neonatais. No entanto, a mortalidade específica para o peso à nascença poderá não ser suficiente como indicador do desempenho das unidades de cuidados intensivos neonatais porque não tem em conta outras diferenças no risco, nomeadamente aquelas que dizem respeito à severidade inicial da doença [64].

O *CRIB* (*Clinical Risk Index for Babies*), foi desenvolvido e validado entre 1988 e 1990. Trata-se de um índice de gravidade clínica para recém-nascidos de muito baixo peso (inferior a 1500 g), determinado pela associação de seis variáveis, como mencionado na secção 6.1. Essas variáveis são o peso à nascença, a idade gestacional, a malformação congénita, o máximo excesso de base nas primeiras 12 horas pós parto, os níveis máximos e mínimos de FiO_2 nas primeiras 12 horas pós parto.

Em [64] faz-se uma comparação dos cuidados oferecidos por unidades de cuidados intensivos neonatais de vários hospitais (hospital 1 - H_1, hospital 2 - H_2, hospital 3 - H_3 e hospital 4 - H_4) usando o *CRIB* como medida de risco neonatal inicial.

Como verificado na secção 6.1, devido à complexidade de recolha de variáveis, e pelo valor da área abaixo da curva ROC, o *CRIB* foi considerado como o melhor indicador do risco de mortalidade neonatal.

Para os recém-nascidos sobrevivos, o índice *CRIB* pode estar associado ao aparecimento de algumas sequelas. *HIVPPVV* (*Imagens ecográficas neurológicas alteradas*), *ROP* (*Retinopatia da Prematuridade*) e *DBP* (*Displasia Bronco-Pulmonar*) são três importantes sequelas que poderão exibir alguma associação com este índice.

Assim para estudar de que forma poderão estar associados o aparecimento de sequelas ao índice *CRIB*, mediu-se esta possível associação utilizando a análise através de curvas ROC e também, a regressão logística, para cada uma das sequelas mencionadas.

6.3.1 Metodologia

Análise ROC

Uma aproximação possível para testar se a diferença entre duas curvas ROC, associadas a conjuntos de dados independentes é significativa, envolve o índice área, A, que sumaria cada curva ROC em termos da área abaixo desta. Aqui, a hipótese nula relevante é que os dois conjuntos de dados em questão, provêm de curvas ROC com áreas abaixo destas semelhantes:

$$H_0 : A_2 - A_1 = 0$$

$$H_1 : A_2 - A_1 \neq 0.$$

Um método para testar se as diferenças entre duas áreas abaixo das curvas *ROC* provenientes de amostras independentes são significativas, consiste na utilização da razão crítica z, definida em 4.10.1.

Regressão logística com variáveis independentes policotómicas

Em qualquer problema de regressão a quantidade chave é o valor médio da variável resposta, dado o valor da variável independente. Esta quantidade é normalmente designada por *média condicional* e pode ser expressa por $E(Y \mid x)$, onde Y designa a variável resposta e x designa o valor da variável independente.

Na regressão linear assume-se que esta média pode ser expressa como uma equação linear em x, do tipo:

$$E(Y \mid x) = \beta_0 + \beta_1 x.$$

A partir desta expressão, verifica-se que $E(Y \mid x)$ pode tomar qualquer valor com x a variar de $-\infty$ a $+\infty$. Contudo, com dados dicotómicos, esta medida deverá estar compreendida entre zero e um, inclusive.

Considere-se $\pi(x) = E(Y \mid x)$. A forma específica do modelo de regressão logística para uma variável resposta dicotómica, tem a forma:

$$\pi(x) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)}. \tag{6.1}$$

A transformação de $\pi(x)$ é denominada por *transformação logit*. Esta transformação é definida em termos de $\pi(x)$, como sendo

$$g(x) = \ln\left(\frac{\pi(x)}{1-\pi(x)}\right)$$
$$= \beta_0 + \beta_1 x.$$

A importância desta transformação é que $g(x)$ tem muitas propriedades desejáveis dos modelos de regressão linear. O *logit* $g(x)$ é linear nos seus parâmetros, é uma função contínua, e pode variar de $-\infty$ a $+\infty$, consoante o domínio de variação de x.

Uma diferença importante entre os modelos de regressão linear e o da regressão logística diz respeito à distribuição condicional da variável resposta. Na regressão linear assume-se que uma observação da variável resposta pode ser expressa como $y = E(Y \mid x) + \varepsilon$, em que ε é designado por *erro*, e dá o desvio de uma observação em relação à média condicional. A hipótese mais comum é que este *erro* ε segue uma distribuição Normal com média zero e variância constante ao longo dos níveis da variável independente; assim resulta que a distribuição da variável resposta dado x, será Normal com média $E(Y \mid x)$, e variância constante. Quando a variável resposta é dicotómica, este pressuposto não se verifica. Nesta situação, deve-se expressar o valor da variável resposta dado x como $y = \pi(x) + \varepsilon$. Aqui a quantidade ε pode assumir um dos dois valores possíveis:

$$Y = 1 \qquad \Longrightarrow \qquad \varepsilon = 1 - \pi(x)$$

com probabilidade $\pi(x)$, e

$$Y = 0 \qquad \Longrightarrow \qquad \varepsilon = -\pi(x)$$

com probabilidade $1 - \pi(x)$. Então, ε tem uma distribuição com média zero e variância igual a $\pi(x)[1 - \pi(x)]$, isto é, a variável resposta segue uma

distribuição binomial com probabilidade dada pela média condicional, $\pi(x)$.

Ajuste do modelo de Regressão Logística

Considere-se uma amostra de n observações independentes do par (x_i, y_i) com $i = 1, 2, \cdots, n$, e y_i e x_i, designam, respectivamente, o valor da variável resposta e o valor da variável independente, correspondente ao *iésimo* indivíduo.

Para ajustar um modelo de regressão logística do tipo do da equação (6.1), torna-se necessário estimar os parâmetros desconhecidos, β_0 e β_1.

Na regressão linear o método mais utilizado é o dos *mínimos quadrados*. Neste método escolhe-se os valores de β_0 e β_1 que minimizam a soma dos quadrados dos desvios dos valores observados de Y em relação aos valores previstos baseados no modelo especificado. Sobre as usuais condições para a regressão linear, o *método dos mínimos quadrados* conduz a estimadores com um número de propriedades estatísticas desejáveis. Infelizmente, quando este método é aplicado a um modelo de resposta dicotómica, os estimadores não apresentam as mesmas propriedades.

O método geral de estimação alternativo ao da função dos mínimos quadrados, para o modelo de regressão linear (quando o termo do *erro* é normalmente distribuído), é o *método da máxima verosimilhança*. Para aplicar este método, tem de se começar por construir a designada *função de verosimilhança*.

A *função de verosimilhança*, expressa a probabilidade dos dados observados como uma função dos parâmetros desconhecidos. Os *estimadores de máxima verosimilhança (EMV)* destes parâmetros, são escolhidos de forma a maximizarem a *função de verosimilhança*.

Para o modelo de regressão logística dicotómica, onde a variável resposta

está codificada por $Y = 0$ e $Y = 1$, a função de probabilidade condicional pode ser expressa através de

$$P(Y \mid x) = \begin{cases} \pi(x) & \text{se } Y = 1 \\ 1 - \pi(x) & \text{se } Y = 0 \end{cases}$$

Assim, para os pares $(x_i,\ y_i)$, quando $y_i = 1$ a contribuição para a *função de verosimilhança* é $\pi(x_i)$, e para os pares cujo valor $y_i = 0$ a contribuição para a *função de verosimilhança* é $1 - \pi(x_i)$, onde a quantidade $\pi(x_i)$ designa o valor de $\pi(x)$ calculada num valor x_i. Uma forma de expressar a contribuição para a *função de verosimilhança* do par $(x_i,\ y_i)$ é através do termo:

$$\zeta(x_i) = \pi(x_i)^{y_i} \left[1 - \pi(x_i)\right]^{1-y_i}.$$

Desde que as observações sejam independentes, a *função de verosimilhança* é dada por:

$$\begin{aligned} l(\beta) &= \prod_{i=1}^{n} \zeta(x_i) \\ &= \prod_{i=1}^{n} \pi(x_i)^{y_i} \left[1 - \pi(x_i)\right]^{1-y_i} \end{aligned} \tag{6.2}$$

onde $\pi(x_i)$ representa $P(Y = 1 \mid x_i)$, também designada por *probabilidade de sucesso.*

O método da máxima verosimilhança estabelece que se utiliza para as estimativas de β, os valores que maximizam a expressão (6.2). Matematicamente, torna-se mais fácil trabalhar a expressão do logaritmo da verosimilhança, dada por

$$L(\beta) = \ln\left[l(\beta)\right] = \sum_{i=1}^{n} \left\{ y_i \ln\left[\pi\left(x_i\right)\right] + \left(1 - y_i\right) \ln\left[1 - \pi\left(x_i\right)\right] \right\}. \tag{6.3}$$

Para achar o valor de β que maximiza $L(\beta)$, deriva-se a expressão (6.3) em ordem a cada parâmetro e igualam-se as expressões obtidas a zero, obtendo-se assim as equações de verosimilhança:

$$\frac{\partial L(\beta)}{\partial \beta_0} = \sum_{i=1}^{n} \left[y_i - \pi\left(x_i\right) \right]$$

$$\frac{\partial L(\beta)}{\partial \beta_1} = \sum_{i=1}^{n} x_i \left[y_i - \pi\left(x_i\right) \right].$$

Para a regressão logística dicotómica, as equações de verosimilhança são não lineares em β, o que requer métodos de resolução de equações não lineares do tipo Newton-Raphson.

Interpretação dos coeficientes estimados

Define-se *razão das possibilidades* da variável resposta $Y = 1$ versus a variável resposta cujo valor é $Y = 0$, para valores da covariável $x = a$ versus $x = b$, como:

$$\Psi\left(a, b\right) = \frac{P(Y = 1 \mid x = a)/P(Y = 0 \mid x = a)}{P(Y = 1 \mid x = b)/P(Y = 0 \mid x = b)}.$$

Num modelo com uma única covariável, onde a variável resposta é binária, o coeficiente do declive do *logit* é idêntico ao logaritmo da *razão das possibilidades (odds ratio)*:

$$\widehat{\beta}_1 = \ln\left[\widehat{\Psi}\left(a, b\right)\right] \Longrightarrow \widehat{\Psi}\left(a, b\right) = \exp\left(\widehat{\beta}_1\right).$$

No caso da existência de variáveis independentes policotómicas, é necessário recorrer à criação de variáveis *design*. Hosmer e Lemeshow [43], discutem os vários métodos de criação destas variáveis. Referem que a escolha de um método particular, poderá depender em alguns casos do objectivo da análise e do estágio de desenvolvimento do modelo.

O agrupamento escolhido para a variável *CRIB*, teve em conta os dados referentes ao hospital 4, de tal modo que a forma e o poder discriminatório das curvas ROC seja mantido, isto é, a área abaixo da curva em função do *CRIB* agrupado seja aproximadamente a mesma.

Neste estudo, para o delineamento dos modelos de regressão logística, utilizou-se o agrupamento da variável CRIB em três classes ($1-[0,3]; 2-[4,6]; 3-[7,20]$) e o método designado por *Indicator (1),* que se passa a descrever.

No caso de uma variável policotómica com três classes, $k = 3$, são necessárias duas variáveis de *design*; por exemplo, segundo o método utilizado, ter-se-ía:

			Variáveis $\curvearrowleft$ *design* $\curvearrowright$	
CRIB	**CRIBAG**	(classe)	D_1	D_2
Baixo	[0 ; 3]	(1)	0	0
Moderado	[4 ; 6]	(2)	1	0
Elevado	≥ 7	(3)	0	1

Este é o método mais utilizado na regressão logística [43], em que o grupo de referência é aquele em que todas as variáveis *design* são iguais a zero.

Hosmer e Lemeshow [43], demonstraram que o valor dos coeficientes do modelo da regressão logística assim obtido, e o valor do *log odds*, é o mesmo.

6.3.2 Descrição dos dados

A amostra em estudo é constituída por 234 recém-nascidos de muito baixo peso (inferior a 1500 g) provenientes de 4 hospitais em Portugal durante o ano de 1995. A distribuição por hospital é: 77 para o hospital 1, 33 para o hospital 2, 45 para o hospital 3 e 79 para o hospital 4. Destes 234 recém-nascidos, 183 sobreviveram e 51 faleceram. As taxas de mortalidade registadas em cada hospital foram 17% para o hospital 1, 24% para o hospital 2, 31% para o hospital 3 e 20% para o hospital 4. Na tabela 6.6 encontra-se uma descrição das variáveis em estudo.

6.3.3 Resultados Experimentais

Nas figuras 6.13, 6.14, 6.15, 6.16, estão representados os gráficos de distribuição de frequências segundo o *CRIB*, para os recém-nascidos de muito baixo peso que faleceram e para os que sobreviveram, para os diferentes hospitais incluídos no estudo.

Tabela 6.6: Descrição das variáveis em estudo

Variável	**Valores**	**Tipo**	**Designação**
Clinical Risk	0, 1,..., 20	Ordinal	*CRIB*
Index for Babies			
CRIB agrupado	1 - $[0, 3]$	Ordinal	*CRIBAG*
em classes	2 - $[4, 6]$		
	3 - ≥ 7		
Hospitais	1	Nominal	H_1
	2		H_2
	3		H_3
	4		H_4
Sequelas			
Imagem ecográfica	0 = Não	Nominal	*HIVPPVV*
Neurológica	1 = menos grave		
	2 = mais grave		
HIVPPVV agrupado	0 = Não	Nominal	*HIVPPVVA*
segundo a existência	1 = Sim		
Retinopatia da	0 = Não	Nominal	*ROP*
Prematuridade	1 = Sim		
Displasia	0 = Não	Nominal	*DBP*
Broncopulmonar	1 = Sim		
Morte	0 = sobrevivo	Nominal	*MORTE*
	1 = falecido		

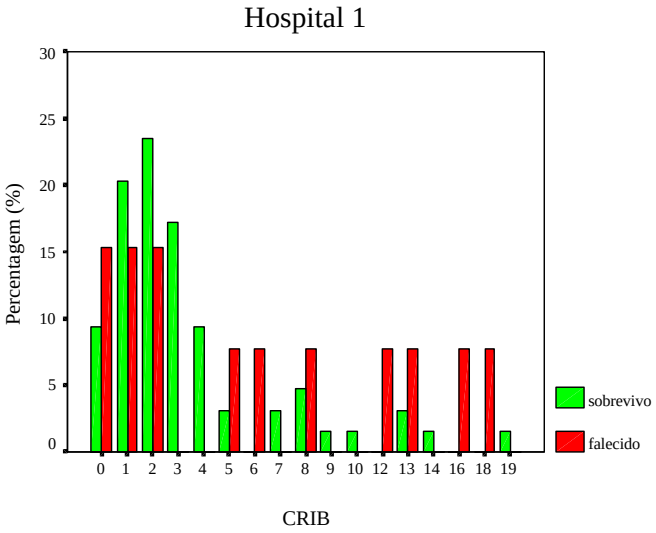


Figura 6.13: Distribuição de frequências segundo o CRIB, para os recém-nascidos de muito baixo peso que faleceram e para os que sobreviveram, para o hospital 1.

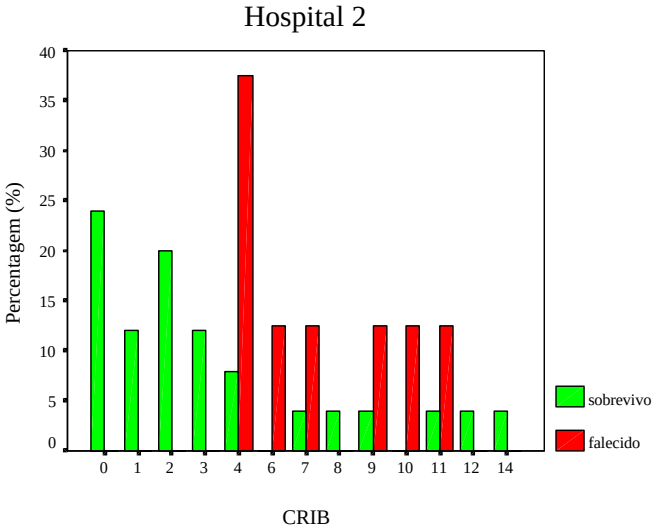


Figura 6.14: Distribuição de frequências segundo o CRIB, para os recém-nascidos de muito baixo peso que faleceram e para os que sobreviveram, para o hospital 2.

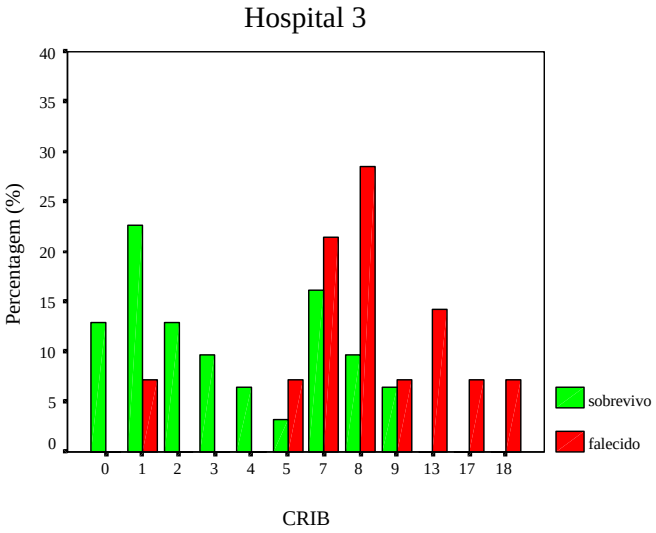


Figura 6.15: Distribuição de frequências segundo o CRIB, para os recém-nascidos de muito baixo peso que faleceram e para os que sobreviveram, para o hospital 3.

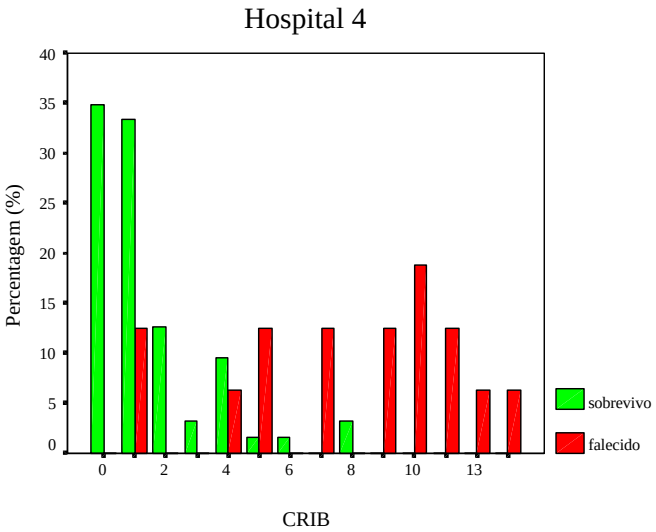


Figura 6.16: Distribuição de frequências segundo o CRIB, para os recém-nascidos de muito baixo peso que faleceram e para os que sobreviveram, para o hospital 4.

Da análise destes gráficos, verifica-se que existe sobreposição das distribuições para os recém-nascidos de baixo peso falecidos e sobrevivos. Note-se ainda, que um aumento na escala de *CRIB* tende a indicar a ocorrência de falecimento.

De notar que, no hospital 1 se verifica a ocorrência de um maior número de mortes, para os valores mais baixos da escala de *CRIB*.

Áreas abaixo da curva ROC e erros padrão

Tal como nas secções anteriores, a metodologia utilizada para o cálculo das áreas abaixo das curvas ROC (A), e respectivos erros padrão ($SE(A)$), foi a aproximação não paramétrica à estatística de Wilcoxon-Mann-Whitney (equação (4.31)) sugerida por Hanley e McNeil [37]. Os resultados encontram-se resumidos nas tabelas 6.7 e 6.8, respectivamente.

Tabela 6.7: Valores de A e $SE(A)$ para os diferentes hospitais na previsão do falecimento segundo a escala do CRIB, para recém-nascidos de muito baixo peso (< 1500 g).

Hospital	**Área abaixo da curva ROC** (A)	**Erro padrão** ($SE(A)$)
H_1	0.59	0.09
H_2	0.79	0.10
H_3	0.84	0.07
H_4	0.92	0.05

Tabela 6.8: Valores de A e $SE(A)$ para as diferentes sequelas segundo a escala do CRIB.

Sequela	**Área abaixo da curva ROC** (A)	**Erro padrão** ($SE(A)$)
HIVPPVVA	0.78	0.05
ROP	0.82	0.05
DBP	0.76	0.07

Modelos de regressão logística com covariável policotómica

Para a construção dos modelos de regressão logística, utilizou-se como covariável, para os três modelos traçados, o *CRIBAG* com as categorias 1, 2 e 3. A metodologia utilizada foi a descrita em 6.3.1. Os três modelos de regressão logística obtidos, encontram-se resumidos na tabela 6.9. Nesta tabela, representa-se para cada caso, a seguinte informação:

(1) as estimativas para os coeficientes, $\widehat{\beta}$;

(2) a estimativa do erro padrão para o coeficiente estimado, $SE(\widehat{\beta})$;

(3) o valor da estatística de Wald, $W = \widehat{\beta}^2/VAR(\widehat{\beta})$;

(4) a estimativa da razão das possibilidades, $\widehat{\Psi}$;

(5) o *valor de prova* do teste de significância dos parâmetros estimados

Da análise destes três modelos, verifica-se que o único coeficiente para o qual não se rejeita a hipótese deste ser nulo, é o *CRIBAG(2)* para a sequela *DBP*.

Tabela 6.9: Modelos de regressão logística univariados, com a covariável CRIBAG para as 3 sequelas.

Sequela	Covariável	$\widehat{\beta}$	$SE(\widehat{\beta})$	W	$\widehat{\Psi}$	*valor p*
$HIVPPVVA$	$CRIBAG(1)$	1.435	0.519	7.64	4.2	0.000
	$CRIBAG(2)$	2.269	0.458	24.56	9.7	0.000
	$CONST.$	-1.920	0.260	54.68		0.000
ROP	$CRIBAG(1)$	1.341	0.617	4.72	3.8	0.030
	$CRIBAG(2)$	2.283	0.533	18.38	9.8	0.000
	$CONST.$	-2.370	0.331	51.38		0.000
DBP	$CRIBAG(1)$	1.935	0.654	8.75	6.9	0.003
	$CRIBAG(2)$	0.811	0.725	1.25	2.3	0.263
	$CONST.$	-2.890	0.388	55.40		0.000

Curvas ROC

Na figura 6.17, representam-se as quatro curvas ROC para os diferentes hospitais e na figura 6.18 as três curvas ROC para as diferentes sequelas segundo a escala de *CRIB*, sendo a representação feita sobre o mesmo espaço ROC.

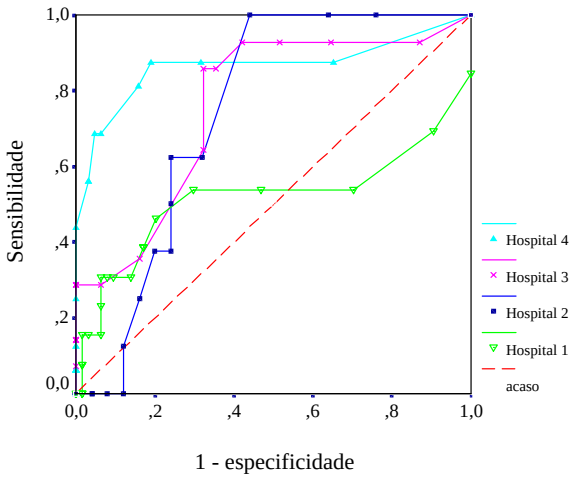


Figura 6.17: Curvas ROC para os 4 hospitais.

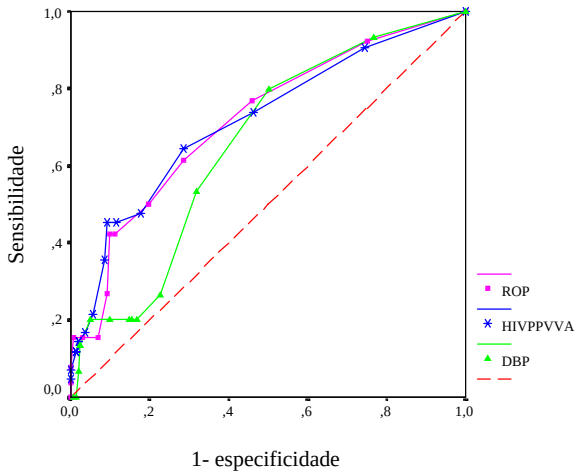


Figura 6.18: Curvas ROC para as 3 sequelas.

Testes de comparação múltipla

Os resultados dos testes de comparação múltipla encontram-se resumidos na tabela 6.10 em termos de *valor de prova*, de acordo com a metodologia referida em 4.10.1, para amostras independentes.

Tabela 6.10: Testes de comparação múltipla entre os diferentes hospitais.

valor de prova	H_1	H_2	H_3
H_2	0.124		
H_3	0.049	0.667	
H_4	0.004	0.136	0.267

6.3.4 Discussão dos resultados

Por comparação das áreas abaixo das curvas ROC para os diferentes hospitais pode dizer-se que o hospital H_4 apresenta um melhor desempenho, em termos de *CRIB*, do que todos os outros, pois é aquele que apresenta maior valor de área abaixo da curva ROC ($A = 0.92$) e menor erro padrão ($SE(A) = 0.05$).

Através da análise de comparações múltiplas, para os quatro hospitais, verifica-se que não existem diferenças significativas em termos de desempenho de cuidados intensivos neonatais, entre o hospital H_1 e os hospitais H_2 e H_3, sendo significativa apenas a diferença entre o hospital H_1 e o hospital H_4. Estes testes foram feitos de forma a garantir globalmente um nível de significância $\alpha = 0.05$, pelo que nas comparações múltiplas entre os vários hospitais se usou um nível de significância de $\alpha/6$ para cada comparação individual.

A análise das curvas ROC para as três sequelas, segundo o *CRIB* para os indivíduos sobrevivos, verificou-se que a sequela ROP apresenta maior valor de área ($A = 0.82$) e menor erro padrão ($SE(A) = 0.05$), o que demonstra a utilidade deste índice como indicador do risco de aparecimento destas sequelas.

Interpretação dos coeficientes do modelo de regressão logística:

· *HIVPPVVA* para $CRIBAG(1)$ $\widehat{\beta} = 1.435$, $\widehat{\Psi}(2,1) = 4.2$, significa que à medida que a escala de *CRIB* aumenta, quando se comparam as classes de CRIBAG 2 e 1, a possibilidade do aparecimento desta sequela aumenta cerca de 4.2 vezes.

· *HIVPPVVA* para $CRIBAG(2)$ $\widehat{\beta} = 2.269$, $\widehat{\Psi}\ (3,1) = 9.7$, significa que à medida que a escala de *CRIB* aumenta, quando se comparam as classes de *CRIBAG* 3 e 1, a possibilidade do apa-

recimento desta sequela aumenta cerca de 9.7 vezes.

· *ROP* para $CRIBAG(1)$ $\widehat{\beta} = 1.341$, $\widehat{\Psi}(2,1) = 3.8$, significa que à medida que a escala de *CRIB* aumenta, quando se comparam as classes de *CRIBAG 2* e 1, a possibilidade do aparecimento desta sequela aumenta cerca de 3.8 vezes.

· *ROP* para $CRIBAG(2)$ $\widehat{\beta} = 2.283$, $\widehat{\Psi}(3,1) = 9.8$, significa que à medida que a escala de *CRIB* aumenta, quando se comparam as classes de *CRIBAG* 3 e 1, a possibilidade do aparecimento desta sequela aumenta cerca de 9.8 vezes.

· Para o *DBP*, verifica-se que a hipótese do coeficiente ser nulo para o $CRIBAG(2)$, não deverá ser rejeitada ($valor\ prova = 0.263$), o que poderá significar, que para este caso os valores intermédios da variável *CRIB* estão fortemente associados ao aparecimento da sequela.

Da análise das sequelas em termos de modelos de regressão logística, verificou-se que de uma forma geral a possibilidade de aparecimento das sequelas aumenta com o aumento da escala de *CRIB*, verificando-se ainda que este aumento se torna mais significativo para as sequelas *HIVPPVVA* e *ROP* para valores de *CRIB* superiores a sete.

Relativamente à sequela *DBP*, verificou-se que o tipo de associação com o *CRIB*, se encontra reflectido no menor valor abaixo da curva ROC determinado para as três sequelas.

Capítulo 7

Programas para o estudo da curva ROC

Os cálculos associados à análise ROC, tal como descritos em capítulos anteriores, são bastante morosos, exigindo cálculo matricial e representações gráficas de alguma complexidade. Naturalmente, foram desenvolvidos alguns programas computacionais para o estudo das curvas ROC. Neste capítulo pretende-se fazer uma exposição sumária, dos principais programas computacionais disponíveis.

O primeiro algoritmo desenvolvido neste campo deve-se a Dorfman e Alf [27] que elaborou um programa em FORTRAN para determinação das estimativas de máxima verosimilhança dos parâmetros de uma curva ROC, considerando o modelo *binormal*.

Uma equipa que desde 1980 tem trabalhado no desenvolvimento de programas para o estudo da curva ROC, liderada por Metz, tem apresentado um conjunto de programas diferentes e específicos para as mais variadas situações, considerando sempre a hipótese do modelo *binormal*.

7.1 ROCFIT

Desenvolvido em Junho de 1989, o programa ROCFIT tem como objectivo estimar, pelo método de máxima verosimilhança, a curva ROC *binormal*, assim como os parâmetros a esta associados, para um conjunto de dados categóricos em classes. A base matemática para este algoritmo de ajuste de uma curva ROC foi desenvolvida por Dorfman em [27], e encontra-se explicada em anexo (Apêndice A).

7.2 LABROC1 e LABROC4

O LABROC1 é um programa para estimação, através do método da máxima verosimilhança, da curva ROC *binormal* e respectivos parâmetros, para conjuntos de dados contínuos. O LABROC4 é a designação para uma versão criada para computadores de grande porte (tipo Workstations). A base matemática para este algoritmo de ajuste de uma curva ROC é a mesma da desenvolvida para o ROCFIT.

Este programa assume que a curva ROC é uma linha recta no plano *binormal*, isto é, no plano cujos eixos coordenados são expressos em termos de "desvios-normais", ou de forma equivalente, assume que os dados originais são provenientes de distribuições normais.

7.3 INDROC

O INDROC foi criado em Junho de 1989, por Charles E. Metz e Helen B. Kronman da Universidade de Chicago [14]. O objectivo deste programa é calcular a significância estatística de diferenças aparentes entre duas curvas ROC no plano *binormal* estimadas a partir de conjuntos de dados categóricos

e independentes, utilizando:

- teste bivariado do qui-quadrado para comparação dos parâmetros estimados a e b das duas curvas ROC;

- teste z univariado, para testar a diferença entre as áreas abaixo das duas curvas ROC (A_z);

- teste z univariado, às fracções de verdadeiros positivos (FVP) de duas curvas ROC num determinado valor de fracção de falsos positivos (FFP_0).

O INDROC assume que os dois conjuntos de dados não são correlacionados, que as categorias para cada conjunto provêm de uma de duas distribuições multinomiais, dos casos designados por *actualmente negativos (normais)*, e dos casos designados por *actualmente positivos (anormais)*.

A aproximação utilizada neste programa, que só é válida para conjuntos de dados independentes, envolve o cálculo das estimativas de máxima verosimilhança dos parâmetros a e b das curvas ROC associados a cada conjunto.

7.4 CORROC

O CORROC, foi um dos primeiros programas a ser desenvolvido pela equipa de Charles E. Metz. Trata-se de um programa específico para dados categóricos correlacionados. Calcula as estimativas de máxima verosimilhança dos parâmetros associados a duas curvas ROC, considerando o modelo *binormal bivariado* como descrito em [60]. Para verificar se a diferença entre duas curvas ROC provenientes de dados categóricos correlacionados é estatisticamente significativa, são utilizados os seguintes testes:

- teste bivariado do qui-quadrado para comparação dos parâmetros estimados a e b das duas curvas ROC;

- teste z univariado, para testar a diferença entre as áreas abaixo das duas curvas ROC (A_z);

- teste z univariado, às fracções de verdadeiros positivos (FVP) de duas curvas ROC num determinado valor de fracção de falsos positivos (FFP_0);

Este programa foi inicialmente desenvolvido para DOS, e posteriormente revisto por Helen B. Kronman, Pu-Lan Wang e Jong-Her Shen em 1980 [16].

7.5 CORROC2

A versão para IBM-PC do CORROC, foi posteriormente desenvolvida pela equipa da Universidade de Chicago, composta por Charles E. Metz, Helen B. Kronman, Pu-Lan Wang, Jong-Her Shen e Ben Herman, em 1993 [16]. A designação para esta nova versão foi CORROC2.

O objectivo deste programa é calcular as estimativas de máxima verosimilhança dos parâmetros para dados ROC classificados em classes e correlacionados, baseando-se no pressuposto de uma distribuição normal bivariada.

Para verificar se a diferença entre duas curvas ROC provenientes de dados classificados em classes e correlacionados, é estatisticamente significativa, são utilizados os mesmos testes do programa CORROC.

Em termos técnicos, o CORROC2 difere do CORROC na medida em que o CORROC2 cria automaticamente os dois conjuntos de matrizes de dados que são necessárias para o algoritmo de estimação da máxima verosimilhança [16].

O CORROC2 tal como o CORROC, utiliza uma modificação do programa de Dorfman (RSCORE II), para obter as estimativas de máxima verosimilhança para os parâmetros a e b e para os limites das classes separadamente para cada curva. Conjuntamente com o cálculo dos coeficientes de correlação directamente a partir das duas matrizes dos dados, estas estimativas são posteriormente utilizadas como pontos iniciais, no método de *scoring* [26], para determinar as estimativas de máxima verosimilhança dos parâmetros considerando o modelo *binormal-bivariado*, assumido para os dados correlacionados, proposto por Metz [60]. Quer o CORROC, quer o CORROC2, são programas escritos em FORTRAN.

7.6 CLABROC

A versão para IBM-PC do CLABROC, foi também desenvolvida pela equipa da Universidade de Chicago, em 1993 [12]. O CLABROC é um programa para tratamento de dois conjuntos de dados contínuos correlacionados. Foi desenvolvido a partir do CORROC2.

Os objectivos do CLABROC são:

- calcular as estimativas de máxima verosimilhança dos parâmetros do modelo para dados contínuos correlacionados e a curva ROC *binormal* associada a estes dados;

- determinar a significância estatística da diferença entre duas curvas ROC, estimada pelos três testes estatísticos desenvolvidos para o CORROC2.

Numa primeira etapa, o programa CLABROC categoriza automaticamente os dados contínuos de uma forma arbitrária, de forma que resultem

no máximo dez classes [12]. De seguida, os conjuntos de dados marginais criados são analisados independentemente através de um programa modificado de Dorfman, para obter as estimativas de máxima verosimilhança dos parâmetros a e b, e para os limites das classes separadamente para cada curva. Da mesma forma do procedimento desenvolvido no CORROC2, estas estimativas vão ser utilizadas como pontos iniciais, no método de *scoring* [26], para determinar as estimativas de máxima verosimilhança dos parâmetros, considerando o modelo *binormal-bivariado*, assumido para os dados correlacionados.

7.7 ROCPWRPC

O ROCPWRPC foi criado em Junho de 1989, pela mesma equipa da Universidade de Chicago. O objectivo do programa ROCPWRPC é prever a potência estatística dos três testes desenvolvidos, quer para amostras independentes quer para amostras correlacionadas, para averiguar diferenças significativas entre duas curvas ROC. O modelo *binormal-bivariado* no qual este programa é baseado encontra-se descrito em [58].

O programa ROCPWRPC, necessita que o utilizador especifique:

(1) os parâmetros a e b assumidos para cada uma das curvas ROC a serem testadas, e o número de categorias utilizadas para definir essas curvas ROC;

(2) os coeficientes de correlação assumidos pelo modelo *binormal-bivariado* para os casos designados por *actualmente negativos* e para os casos designados por *actualmente positivos* segundo as distribuições na variável de decisão;

(3) a razão entre o número de casos *actualmente positivos* (*anormais*) e o número de casos *actualmente negativos* (*normais*), o qual se assume igual para os dois conjuntos de dados;

(4) conjuntos de fracções de falsos positivos, correspondentes aos valores esperados dos pontos de operação assumidos nas duas curvas ROC.

O programa calcula então as variâncias e covariâncias para as estimativas dos parâmetros a e b para as duas curvas ROC utilizando os dados esperados associados aos limites das categorias da variável de decisão (*valores de corte*) que são resultado das fracções de falsos positivos especificadas. Finalmente, o programa utiliza as variâncias e covariâncias calculadas para determinar a potência estatística dos três testes como função do número de casos *actualmente negativos* [15]. Os testes estatísticos são:

(1) *teste bivariado do qui-quadrado* de diferenças simultâneas entre os parâmetros a e b das duas curvas ROC; A potência é calculada em termos de uma distribuição do χ^2 não centrado com dois graus de liberdade;

(2) *teste z univariado*, para testar a diferença entre as áreas abaixo das duas curvas ROC (A_z); A potência é calculada em termos de uma distribuição Normal padrão não-centrada, com a variância da diferença do índice área aproximada em termos da variância e covariância dos parâmetros ROC, por uma expressão convencional da derivada parcial de primeira ordem [15].

(3) *teste z univariado (bilateral), às fracções de verdadeiros positivos* (FVP) de duas curvas ROC num determinado valor de fracção de falsos positivos (FFP_0); A potência é calculada para

valores de fracção de falsos positivos iguais a 0.02, 0.05, 0.10, 0.15, 0.20 e 0.25 essencialmente da mesma forma utilizada para o teste anterior.

7.8 LABMRMC

O algoritmo empregue no programa LABMRMC foi delineado por Donald Dorfman, Kevin Berbaum e Charles E. Metz, e foi escrito por Benjamin A. Herman e Hatem AbuDagga, e encontra-se disponível numa versão beta para PC, desde Abril 1997.

Os objectivos do programa LABMRMC são:

- calcular as estimativas dos parâmetros do modelo *binormal bivariado* para dados contínuos ou discretos, em classes, até cinco potenciais testes de diagnóstico correlacionados (*tratamentos*, na terminologia estatística) e até *rmax* leitores de imagem (*rmax* é cerca de dez nesta versão, mas este número pode ser configurado pelo utilizador) e assim, estimar as curvas ROC *binormais* associadas a estes dados;

- calcular a significância estatística da diferença entre as médias das áreas abaixo das curvas ROC que são estimadas para os dois testes de diagnóstico (isto é, tratamentos), utilizando a metodologia *jackkinife* e ANOVA [25].

O LABMRMC assume que a verdadeira curva ROC para cada combinação leitor-tratamento é representada por uma linha recta nos eixos coordenados de *desvios normais*.

Num primeiro passo, LABMRMC categoriza automaticamente os dados contínuos no sentido de produzir uma gama apropriada de pontos de

operação em cada curva ROC. Os conjuntos de dados marginais criados desta forma, são então analisados, independentemente, para obter as estimativas de máxima verosimilhança dos parâmetros convencionais, a *ordenada na origem*, a, o *declive*, b, e os *limites das classes* separadamente para cada curva ROC *binormal.*

De seguida, o programa LABMRMC a metodologia *jackkinife* para determinação das estimativas dos designados *pseudovalores* do índice área abaixo da curva ROC, A_z, e com estes através da metodologia ANOVA, determinar a significância estatística entre as condições.

7.9 ROCKIT

O programa ROCKIT é a combinação dum conjunto de programas desenvolvidos pela equipa da Universidade de Chicago, num único programa com capacidades adicionais, como por exemplo, analisar conjuntos de dados parcialmente correlacionados. Este programa substitui o ROCKFIT, LABROC1, INDROC, CORROC2, e CLABROC. O ROCKIT está delineado para ajustar curvas ROC *binormais* quer para conjuntos de resultados de diagnóstico contínuos, quer categóricos.

Os objectivos do ROCKIT são:

- calcular as estimativas de máxima verosimilhança dos parâmetros do modelo convencional *binormal* para os dados introduzidos;

- calcular as estimativas de máxima verosimilhança dos parâmetros do modelo *binormal bivariado* para dados de dois testes de diagnóstico potencialmente correlacionados, e também estimar as curvas ROC binormais resultantes destes dados e as suas correlações;

- determinar a significância estatística da diferença entre duas curvas ROC, estimada por um dos três testes estatísticos desenvolvidos para INDROC ou CORROC2.

Neste programa são permitidos dados de três tipos:

(1) resultados de testes não emparelhados. As duas condições são aplicadas a amostras independentes - por exemplo, dois testes de diagnóstico diferentes aplicados a pacientes diferentes, ou dois radiologistas que fazem os seus juízos no que diz respeito à presença de uma determinada doença em imagens diferentes;

(2) resultados de testes correlacionados na totalidade, nos quais os dados para as duas condições são medidos para cada caso numa única amostra. Às duas condições para cada par teste-resultado poderão corresponder, por exemplo, dois testes de diagnóstico diferentes realizados no mesmo paciente, ou dois radiologistas que fazem os seus juízos no que diz respeito à presença de uma determinada doença na mesma imagem;

(3) resultados parcialmente correlacionados - por exemplo, dois testes de diagnóstico diferentes efectuados na mesma amostra de pacientes e que para um conjunto adicional de pacientes, apenas é efectuado um teste de diagnóstico.

O ROCKIT assume que a verdadeira curva ROC para cada condição é representada por uma linha recta no plano *binormal*.

Tal como no LABMRMC, o ROCKIT começa por categorizar automaticamente os dados contínuos no sentido de produzir uma gama apropriada de pontos de operação em cada curva ROC. Os conjuntos de dados marginais criados desta forma são então analisados, para obter as estimativas de

máxima verosimilhança dos parâmetros, *a* (*ordenada na origem)*, *b* (*declive)*, e *limites das classes* separadamente para cada curva ROC *binormal.*

Se os dados são correlacionados ou parcialmente correlacionados, então os coeficientes de correlação são calculados directamente das matrizes de dados bivariados categóricos para os casos *actualmente positivos* e *actualmente negativos*, e posteriormente utilizadas pelo ROCKIT como estimativas iniciais, para calcular (através do método de *scoring*) as estimativas de máxima verosimilhança dos parâmetros do modelo *binormal bivariado.* O procedimento matemático aplicado pelo ROCKIT, encontra-se desenvolvido em [60].

7.10 AccuROC

A primeira versão do programa AccuROC foi criada para DOS em 1993 por Vida [85]. Em Janeiro de 1999 é apresentada a versão 1.2 para Windows 95 e, mais tarde, surge a versão 2.0 para Windows 95/98/NT. A versão 1.2 do AccuROC é uma versão mais evoluída do que a versão inicial descrita por Vida em [85], com a adição de um interface gráfico e novas capacidades estatísticas.

O AccuROC utiliza métodos não paramétricos como as estatísticas de Mann-Whitney e do qui-quadrado para calcular e representar curvas ROC para amostras individuais e compara curvas ROC para amostras independentes e para amostras correlacionadas.

A versão 2.0 do AccuROC para Windows:

- efectua a análise ROC para um único teste, utilizando o método de DeLong;
- compara curvas ROC para amostras independentes utilizando o erro padrão desenvolvido por DeLong;

- compara duas ou três curvas ROC para amostras correlacionadas utilizando a metodologia desenvolvida por DeLong;
- utiliza a metodologia de *bootstrap*, para determinar erros padrão, limites de confiança, percentis, correcção de tendência e limites de confiança acelerados, através da metodologia desenvolvida por Efron e Tbishirani;
- representa graficamente uma, duas ou três curvas ROC;
- permite copiar e colar as curvas ROC para outros documentos do Windows 95;
- permite exportar as coordenadas dos gráficos para outros pacotes gráficos;
- efectua os cálculos para determinar a potência estatística permitindo estimar a dimensão da amostra requerida;
- calcula uma grande variedade de medidas de desempenho do teste para cada valor de corte.

7.11 Outros

Existem pacotes estatísticos de carácter mais generalizado, como por exemplo o *S-Plus* e o *SPSS*, que nas suas últimas versões incluem a análise da curva ROC, apenas para um conjunto de dados. Efectuam o cálculo da área abaixo da curva ROC e respectivo erro padrão pelo método não paramétrico da estatística de Wilcoxon-Mann-Whitney, determinam os valores das coordenadas do gráfico ROC no plano ROC unitário e traçam a curva ROC empírica. No caso do *SPSS*, é também dado como opção o ajuste segundo o modelo *bi-exponencial.*

Capítulo 8

Novo programa - ROCNPA

8.1 Motivação

Como descrito no capítulo 7, existe uma vasta gama de programas desenvolvidos nos mais variados campos da análise ROC, desde o ajuste da curva ROC, passando pela determinação de medidas de diagnóstico como a sensibilidade e especificidade, à avaliação de desempenho de sistemas de diagnóstico.

A grande maioria dos programas desenvolvidos baseia-se numa abordagem paramétrica, considerando modelos como o *binormal* ou o *binormal bivariado* (no caso de duas amostras correlacionadas).

Por outro lado, o único programa que efectua uma abordagem não paramétrica, o AccuROC, fá-lo no máximo para um conjunto de três testes de diagnóstico. Saliente-se ainda que em termos gráficos todos os programas desenvolvidos até à data apresentam diversas lacunas. Entre elas, salienta-se a impossibilidade de visualização directa da curva ROC empírica.

Por último, todos estes programas foram desenvolvidos apenas para plataformas de suporte como DOS e WINDOWS. Assim, devido às dificuldades encontradas na análise dos dados trabalhados no capítulo 6, procurou-se

criar um programa que se tornasse versátil para a análise ROC, tendo como objectivos principais:

- determinar um ajuste para a curva ROC;

- avaliar o desempenho do teste de diagnóstico através de um índice de determinação simples e livre de hipóteses distribucionais;

- comparar mais do que três sistemas de diagnóstico, quer os dados sejam provenientes de amostras independentes, quer de amostras correlacionadas.

8.2 Requisitos do ROCNPA

8.2.1 Requisitos do sistema

O ROCNPA foi desenvolvido para poder ser utilizado em qualquer tipo de máquina que possua no mínimo 32 MB de memória RAM. Por outro lado encontra-se preparado para correr em qualquer tipo de sistema operativo, como WINDOWS, LINUX, UNIX e MacOSX.

8.2.2 Notas

A linguagem utilizada na elaboração do ROCNPA foi o JAVA que por ser uma linguagem que tem por base a programação orientada aos objectos, apresenta um conjunto de requisitos bastante atractivos. O JAVA corre sobre um ambiente específico denominado *máquina virtual*. Assim, o seu maior atractivo é o facto de se poder correr qualquer programa desenvolvido em JAVA em qualquer plataforma (sistema operativo + CPU) para a qual esteja já desenvolvida a *máquina virtual*.

Neste momento, existe *máquina virtual* de JAVA para WINDOWS 95/ 98/ NT/ 2000 (Intel), UNIX (Solaris), LINUX (Intel) e Macintosh (MacOSX) Por este facto, a máquina onde irá ser instalado o ROCNPA terá de ter instalada a versão 1.2.2 ou superior do JDK (JAVA Development Kit) ou JRE (JAVA Runtime Environment) ambos disponíveis para *download* em *http://java.sun.com/*.

8.3 Linguagem JAVA

A linguagem JAVA começou a ser desenvolvida no início dos anos 90 no seio de uma pequena equipa de engenheiros de software da Sun Microsystems, liderada por James Gosling [52]. O objectivo desta equipa era desenvolver uma linguagem para equipamentos electrónicos com "*chips*" programáveis tais como torradeiras, máquinas de lavar, agendas electrónicas de bolso, entre outros. Os principais requisitos da linguagem a desenvolver, eram a robustez e a segurança (pois os utilizadores de tais dispositivos não admitem erros ou falhas), o baixo custo (os programas teriam de ser simples) e a independência dos "*chips*" (dado que os construtores com grande facilidade os substituem por outros).

O que torna o JAVA uma linguagem muito atractiva, para toda a indústria de computadores, é que não é apenas uma nova linguagem de programação, que é efectuada por objectos e quase totalmente pura (contrariamente ao C++), mas sobretudo porque o JAVA se posicionou como um atractivo e apropriado ambiente de programação e desenvolvimento de aplicações no contexto actual, principalmente a partir do lançamento do sistema JDK (JAVA Development Kit).

A Sun Microsystems apresenta o JAVA como sendo um linguagem sim-

ples, orientada aos objectos, distribuída, interpretada, robusta, segura, neutra em termos de arquitectura, portável, com bom desempenho, de múltiplas "*threads*", dinâmica e orientada para a Internet [52].

O JAVA é considerado simples porque, apesar de ter herdado muitas construções das linguagens C e C++, eliminou um razoável conjunto de construções responsáveis pela "pouca transparência" e grande obscuridade semântica e complexidade dos programas em C e C++, nomeadamente alguns apontadores e alocações de memória entre outros. Por outro lado, o JAVA foi criado como linguagem de programação orientada aos objectos, ao contrário do C++, que consistiu numa quase bem conseguida extensão de C para programação por objectos.

A robustez da linguagem JAVA deve-se ao facto do JAVA ser uma linguagem fortemente tipada. Não tem *apontadores*, todos os acessos a *arrays* e *strings* são validados pelo compilador. As conversões entre tipos são estaticamente verificadas. Possui mecanismos para captura e tratamento de excepções, ou seja, a ocorrência de uma dada excepção durante a execução do programa pode ser tratada através de instruções próprias codificadas pelo programador.

A segurança dos programas tem não só a ver com a garantia de que a sua execução não vai corromper a máquina onde este é executado, mas também com possíveis garantias quanto à sua origem [52]. O JAVA é seguro por ser robusto e também por possuir processos internos de verificação do designado *bytecode* (código interpretado pela máquina virtual), que é o código gerado pelo compilador. Por outro lado, o JAVA permite incluir chaves criptográficas no próprio código, possibilitando deste modo a identificação da origem do mesmo.

O JAVA é ainda uma linguagem cujo desempenho não pode ser compa-

rado com o conseguido por linguagens como o C ou mesmo o Pascal para certas aplicações. Porém, sendo uma linguagem distribuída e orientada para a Internet, a sua eventual falta de desempenho é relativizada se comparada com as velocidades de comunicação e transmissão de dados. Possibilita ainda a execução simultânea de processos livres que realizam diferentes tarefas, pelo que é considerada uma linguagem de múltiplas "*threads*".

Devido a este vasto conjunto de qualidades da linguagem JAVA, optou-se pela sua utilização na realização do programa para estudo das curvas ROC.

8.4 Descrição do ROCNPA

8.4.1 Introdução dos dados

Os dados poderão ser introduzidos de três formas diferentes:

- directamente a partir do teclado;

- a partir de um ficheiro de dados do EXCEL, por simples cópia dos valores da variável;

- a partir de um ficheiro previamente criado.

A janela de diálogo, aquando da iniciação do programa, é do tipo da apresentada na figura 8.1.

Se se optar pela criação de um novo ficheiro de dados, o programa apresenta uma janela de comando para questionar quantas variáveis estão em estudo, como exemplificado na figura 8.2. Caso o número de variáveis seja superior ou igual a dois, ter-se-á logo de caracterizar a amostra, isto é, identificar se se trata de dados provenientes de amostras correlacionadas ou independentes.

Após a caracterização da amostra, o utilizador depara-se com um conjunto de janelas com caixas de diálogo muito simples, que lhe permitem completar

a definição da sua amostra, ou seja, a atribuição de nome às variáveis, qual o valor que corresponde ao teste positivo (se são os valores menores ou maiores da escala que correspondem ao teste positivo) e por fim como é caracterizada a variável resultado. Todo este procedimento de escolha encontra-se ilustrado nas figuras 8.3, 8.4, 8.5.

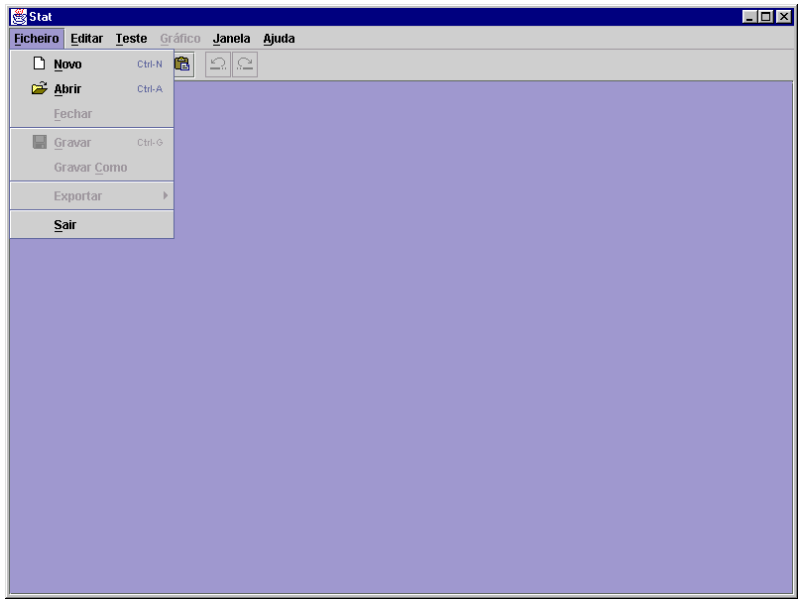


Figura 8.1: Janela do ROCNPA para abrir ou criar um ficheiro de dados.

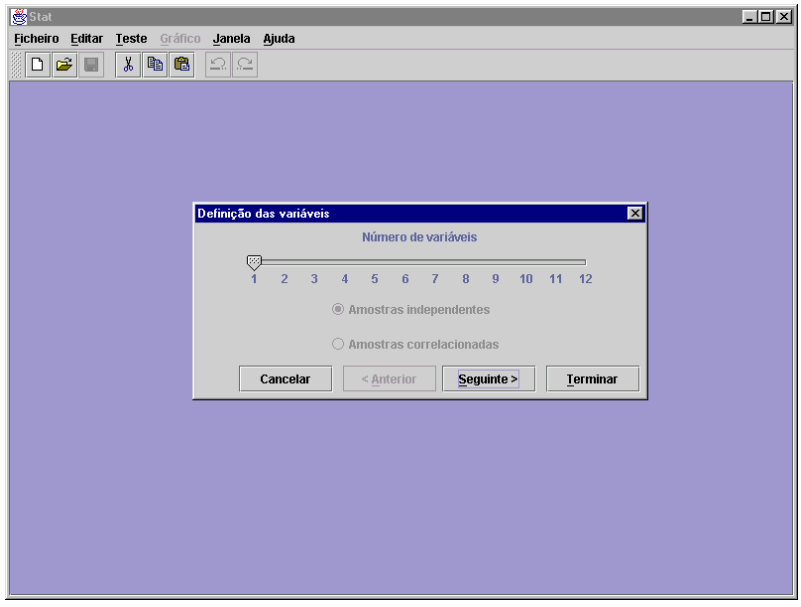


Figura 8.2: Janela de diálogo para caracterização da amostra.

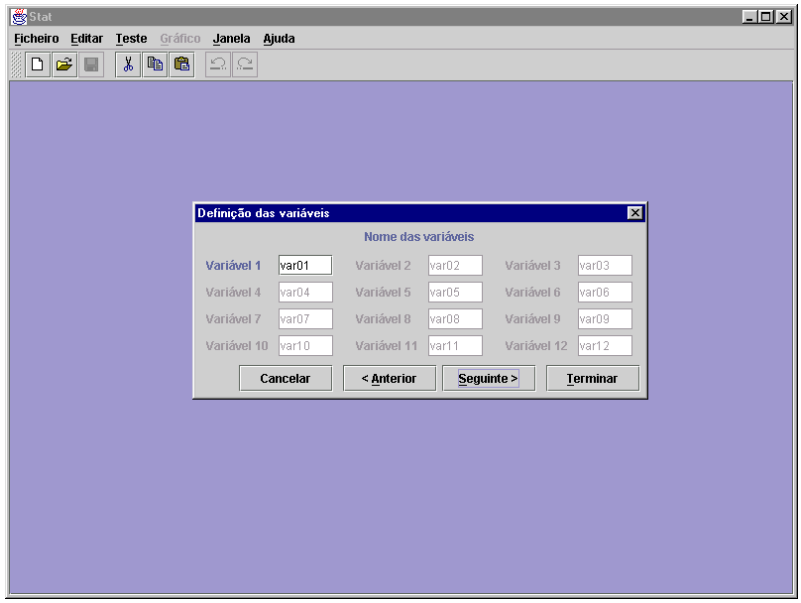


Figura 8.3: Janela de diálogo para a definição dos nomes das variáveis.

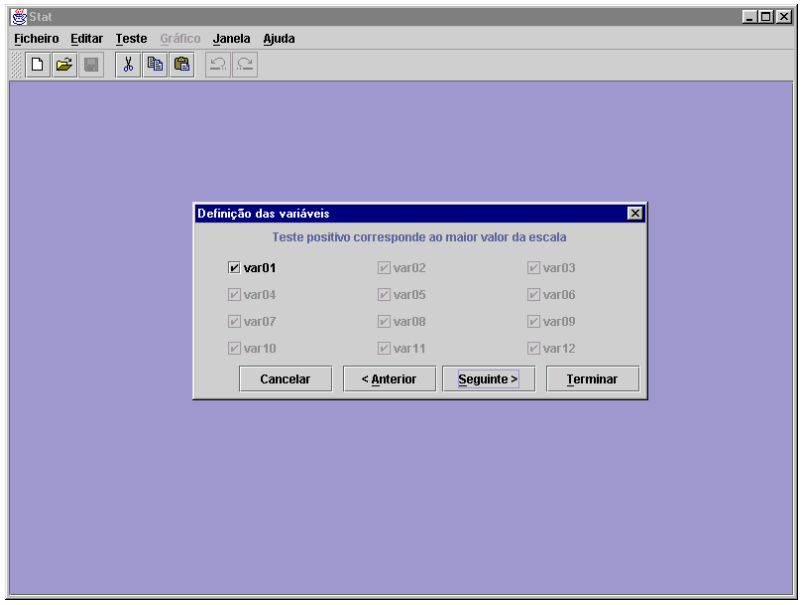


Figura 8.4: Definição das escalas.

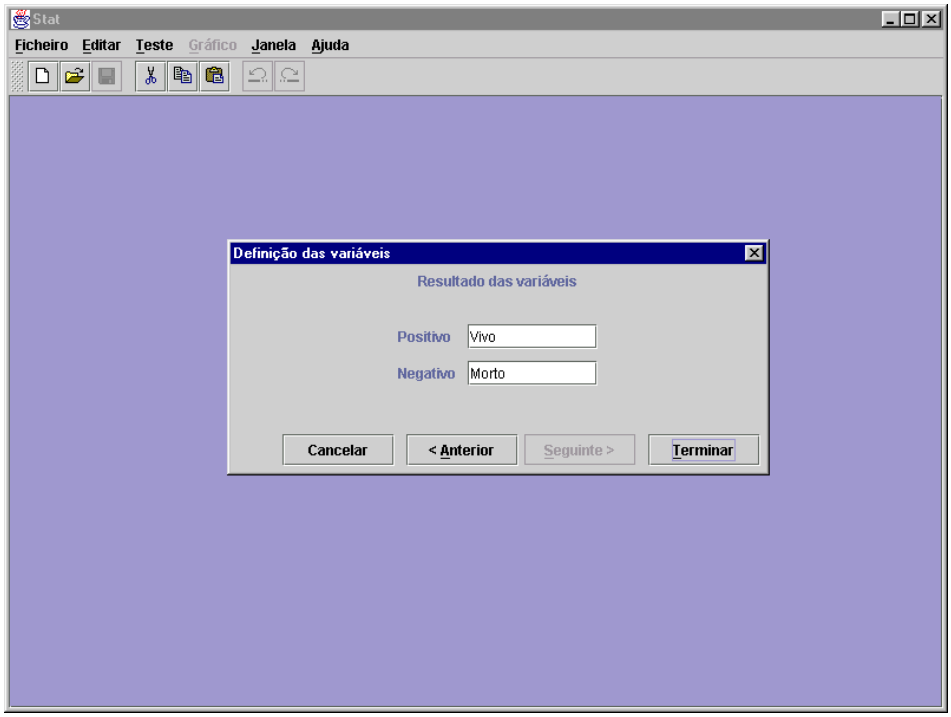
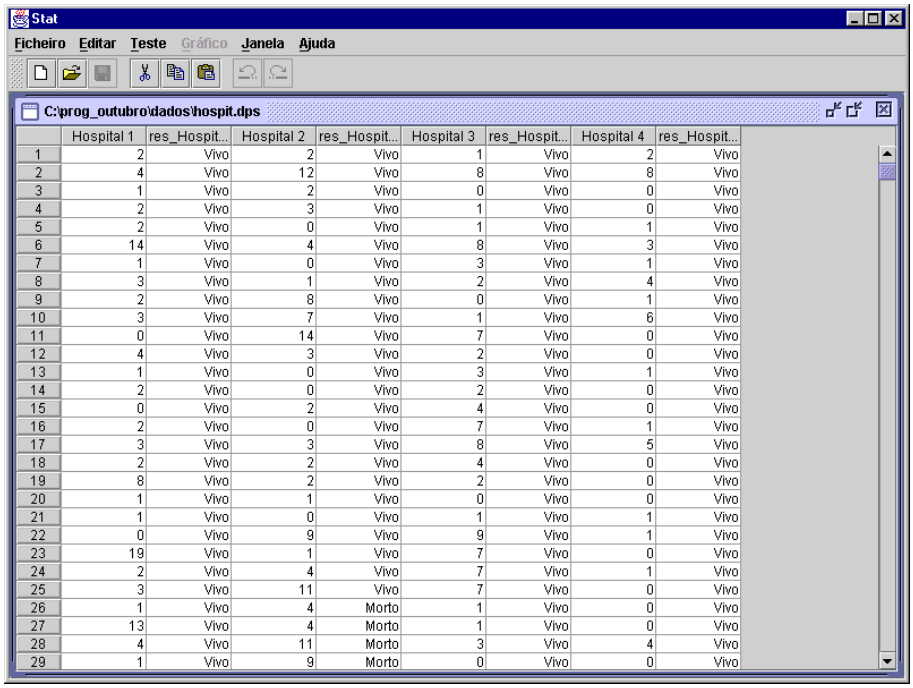


Figura 8.5: Caracterização do resultado.

Na figura 8.6, apresenta-se o aspecto da janela de dados para quatro amostras independentes, cujo maior valor da escala corresponde ao teste positivo e o resultado positivo significa falecimento (morto) e o negativo significa sobrevivência (vivo). Depois de criado o ficheiro de dados, ele pode ser guardado com um nome e uma extensão *.roc* para posterior utilização.



	Hospital 1	res_Hospit...	Hospital 2	res_Hospit...	Hospital 3	res_Hospit...	Hospital 4	res_Hospit...
1	2	Vivo	2	Vivo	1	Vivo	2	Vivo
2	4	Vivo	12	Vivo	8	Vivo	8	Vivo
3	1	Vivo	2	Vivo	0	Vivo	0	Vivo
4	2	Vivo	3	Vivo	1	Vivo	0	Vivo
5	2	Vivo	0	Vivo	1	Vivo	1	Vivo
6	14	Vivo	4	Vivo	8	Vivo	3	Vivo
7	1	Vivo	0	Vivo	3	Vivo	1	Vivo
8	3	Vivo	1	Vivo	2	Vivo	4	Vivo
9	2	Vivo	8	Vivo	0	Vivo	1	Vivo
10	3	Vivo	7	Vivo	1	Vivo	6	Vivo
11	0	Vivo	14	Vivo	7	Vivo	0	Vivo
12	4	Vivo	3	Vivo	2	Vivo	0	Vivo
13	1	Vivo	0	Vivo	3	Vivo	1	Vivo
14	2	Vivo	0	Vivo	2	Vivo	0	Vivo
15	0	Vivo	2	Vivo	4	Vivo	0	Vivo
16	2	Vivo	0	Vivo	7	Vivo	1	Vivo
17	3	Vivo	3	Vivo	8	Vivo	5	Vivo
18	2	Vivo	2	Vivo	4	Vivo	0	Vivo
19	8	Vivo	2	Vivo	2	Vivo	0	Vivo
20	1	Vivo	1	Vivo	0	Vivo	0	Vivo
21	1	Vivo	0	Vivo	1	Vivo	1	Vivo
22	0	Vivo	9	Vivo	9	Vivo	1	Vivo
23	19	Vivo	1	Vivo	7	Vivo	0	Vivo
24	2	Vivo	4	Vivo	7	Vivo	1	Vivo
25	3	Vivo	11	Vivo	7	Vivo	0	Vivo
26	1	Vivo	4	Morto	1	Vivo	0	Vivo
27	13	Vivo	4	Morto	1	Vivo	0	Vivo
28	4	Vivo	11	Morto	3	Vivo	4	Vivo
29	1	Vivo	9	Morto	0	Vivo	0	Vivo

Figura 8.6: Janela de dados no ROCNPA para um conjunto de quatro variáveis independentes.

8.4.2 Análise através de uma curva ROC

Após a introdução dos dados, o ROCNPA permite efectuar a análise através das curvas ROC. Esta análise pode ser efectuada de uma forma completa, através do comando $< Executar\ todos >$ no menu $< Testes >$ da janela de dados, onde o programa produz um conjunto de resultados para análise, nomeadamente os gráficos das distribuições de frequências para cada variável, as curvas ROC empíricas, as curvas ROC no plano *binormal*, as curvas ajustadas no plano unitário, os resultados em termos de área abaixo da curva ROC e respectivos erros padrão e também os valores de prova resultantes dos testes de comparações múltiplas. Ainda no menu $< Testes >$, é permi-

tido executar cada um destes processos individualmente.

Cada um dos resultados mencionados é produzido numa janela em separado, permitindo posteriormente a sua visualização individual ou em simultâneo. Pretende-se que com os gráficos de distribuições de frequências, o utilizador tenha uma ideia do comportamento da variável de um forma simples e rápida. As curvas ROC empíricas são produzidas pela união dos pontos coordenados, que correspondem aos pares $(1-especificidade,\ sensibilidade)$, calculados para cada caso. Posteriormente, estas probabilidades P são transformadas em valores z, que vão constituir as coordenadas de um novo gráfico. O plano assim resultante é designado por plano *binormal*, e a curva ROC representada neste eixo coordenado é a curva ROC *binormal*. Um bom ajuste à curva ROC no plano unitário poderá ser produzido através do ajuste de uma recta no plano *binormal*, salvaguardando as hipóteses subjacentes a este modelo, como descrito no capítulo 2.

O índice área abaixo da curva ROC é determinado por três processos diferentes: a regra do trapézio, a aproximação não paramétrica à estatística de Wilcoxon-Mann-Whitney e a aproximação no plano *binormal* através dos coeficientes estimados da recta de regressão. Os valores dos erros padrão são determinados pela rotina sugerida por Hanley e McNeil [37].

Na figura 8.7 encontra-se exemplificado o conjunto de janelas de resultados produzidas para uma única curva ROC.

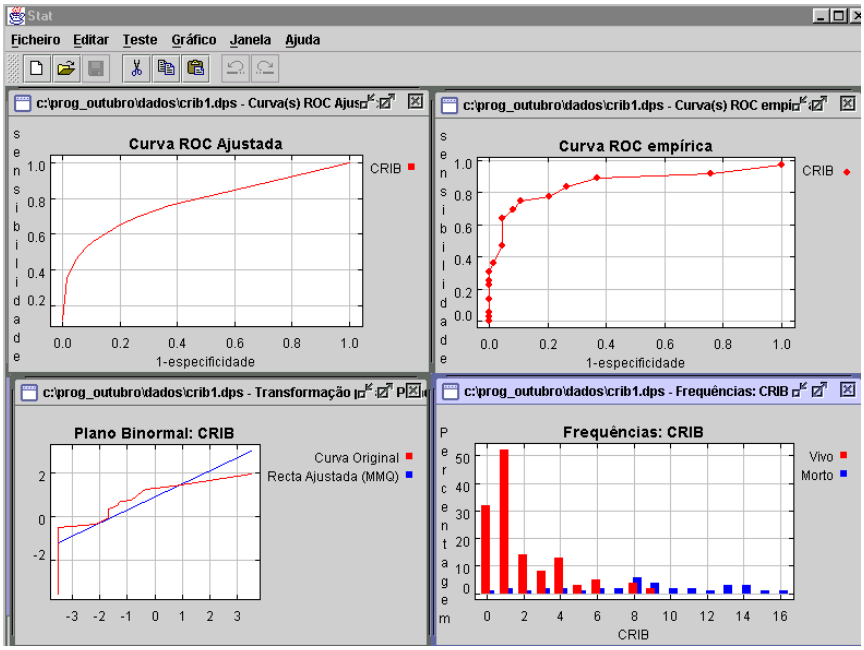


Figura 8.7: Janela de gráficos produzidos no estudo de um único conjunto de dados.

8.4.3 Comparação de duas ou mais curvas ROC

O ROCNPA permite a comparação de mais do que três curvas ROC quer se trate de dados provenientes de amostras independentes ou correlacionadas. Tal facto supera os programas existentes nesta área. A comparação é efectuada através do índice área abaixo da curva ROC por comparações múltiplas dois a dois, utilizando a estatística de teste z, definida por Hanley e McNeil [38].

Para os casos de dados provenientes de amostras correlacionadas, o coeficiente de correlação entre áreas é calculado utilizando o procedimento desenvolvido por DeLong e DeLong [22].

Os resultados produzidos em termos gráficos são as distribuições de frequências para cada variável, as curvas ROC empíricas, as curvas ROC no plano *binormal*, as curvas ajustadas no plano unitário. Os resultados em termos

analíticos são os valores da área abaixo da curva ROC e respectivos erros padrão e também os valores de prova resultantes dos testes de comparações múltiplas.

8.5 Comparação de programas para a curva ROC

Nesta secção procurar-se-à analisar alguns dos programas disponíveis para análise de dados através de curvas ROC, utilizando para cada caso específico os dados tratados no capítulo 6.

Para análise de uma única curva ROC, optou-se por utilizar o SPSS 9.0 devido à sua versatilidade como pacote estatístico.

Seleccionou-se também, um conjunto de programas desenvolvidos pela equipa de Metz, como por exemplo o LABMRMC, o CLABROC e o ROCKIT, que utilizam uma abordagem paramétrica. Para uma abordagem não paramétrica utiliza-se o AccuROC 2.3 desenvolvido por Vida [86] e o programa desenvolvido no âmbito deste trabalho.

8.5.1 Análise de um conjunto de dados

Quando se pretende analisar apenas um conjunto de dados, no contexto da análise ROC, procura-se numa primeira fase, como em qualquer abordagem de análise de dados, averiguar qual a natureza da variável em estudo.

Com o objectivo de ilustrar o desempenho de cada um dos programas em estudo, utilizou-se a variável CRIB já estudada no capítulo 6. Trata-se uma variável ordinal, cuja escala varia entre 0 e 20, e tal como referido nesse capítulo, trata-se de um índice de risco neonatal inicial. Os dados

dizem respeito a 169 recém-nascidos de muito baixo peso, e foram recolhidos durante o ano de 1995, na Unidade de Cuidados Intensivos Neonatais do Hospital Garcia de Orta. A análise será efectuada pelos programas SPSS, ROCKIT, AccuROC e ROCNPA.

A introdução de dados em programas estatísticos como o SPSS é relativamente simples, pois as novas versões já se encontram preparadas para importação de dados directamente a partir de outros programas, como por exemplo o EXCEL. Por outro lado, a introdução de dados directamente a partir do SPSS é muito fácil, pois a sua folha de dados está preparada para a correcta identificação de variáveis e introdução de dados.

Como em qualquer análise estatística de dados, começou-se por fazer uma análise descritiva da variável CRIB, traçando tabelas de frequências, gráficos de barras que permitissem visualizar a distribuição dos valores para os recém-nascidos sobrevivos e falecidos. De seguida, através do comando $[Graphs] \rightarrow [ROC\ Curve]$ inicia-se o processo de análise através da curva ROC. Após esta selecção de comandos aparece uma nova janela para selecção da variável em estudo e das opções que se pretende no que diz respeito ao gráfico, como o traçar a diagonal, intervalos de confiança e coordenadas dos pontos da curva ROC.

Da análise dos resultados obtidos através do SPSS, verifica-se que este permite efectuar o ajuste a um conjunto particular de dados, apresentando um ajuste para a curva ROC assim como o valor da área abaixo da curva ROC e o respectivo erro padrão, determinados pela aproximação à estatística de Wilcoxon-Mann-Whitney. Produz ainda os limites do intervalo de confiança para um determinado grau de confiança, e os valores de *sensibilidade* e $1 - especificidade$ para um conjunto de valores de corte. Os resultados desta análise encontram-se no anexo C.

Efectuando a análise no programa ROCKIT, verifica-se que a introdução dos dados pode ser feita directamente a partir do teclado, ou então (o que é aconselhado pelos autores), através de ficheiros do WORD ou EXCEL convertidos posteriormente para um formato de texto, como referido em [13]. A criação dos ficheiros de dados a partir do EXCEL ou WORD é aconselhável porque a sua introdução directa através do teclado é fastidiosa e morosa quando se trata de dados não agrupados.

Este programa produz um conjunto de valores resultantes da estimativa de máxima verosimilhança para a curva ROC *binormal*, como se pode ver no anexo C. Não produz, no entanto, qualquer tipo de gráfico, fornecendo as coordenadas para a curva ROC ajustada pelo procedimento paramétrico. Para traçar o gráfico, ter-se-à de recorrer a um programa com módulo gráfico, como por exemplo o EXCEL.

No AccuROC a introdução dos dados tem de ser efectuada também através de um ficheiro de texto com características específicas e cuja extensão é .roc. Os resultados produzidos pelo AccuROC são os valores de *sensibilidade* e $especificidade$, área abaixo da curva ROC e respectivo erro padrão utilizando uma abordagem não paramétrica e intervalos de confiança para o índice área abaixo da curva ROC. Apresenta ainda a significância estatística para a comparação do valor da área abaixo da curva ROC obtido com o valor 0.5 (diagonal não informativa). Em termos gráficos produz a curva ROC empírica no plano unitário.

Em termos de análise de resultados, a tabela 8.1 apresenta o resumo dos valores obtidos em cada um dos programas testados, para o índice área abaixo da curva ROC e respectivos erros padrão, considerando a variável CRIB.

A utilização do subscrito z, em A_z, significa que o valor da área abaixo da curva ROC é determinado a partir da sua forma funcional *binormal,* através

Tabela 8.1: Resumo dos valores obtidos para o índice área abaixo da curva ROC

SPSS	**ROCKIT**	**AccuROC**	**ROCNPA**
$A = 0.899$ $SE(A) = 0.034$	$A = 0.899$ $SE(A) = 0.035$	$A = 0.899$ $SE(A) = 0.034$	$A = 0.899$ $SE(A) = 0.034$
	$A_z = 0.901$ $SE(A_z) = 0.033$		$A_z = 0.901$ $SE(A_z) = 0.033$

da equação (5.12).

Como seria de esperar, o valor de A obtido em cada um dos programas é o mesmo, pois a abordagem não paramétrica utilizada em todos eles é a mesma, isto é, a aproximação à estatística de Wilcoxon-Mann-Whitney.

No gráfico da figura 8.8 procurou-se traçar os ajustes dados pelo SPSS e pelos pontos coordenados fornecidos pelo ROCKIT.

Verifica-se que o ajuste à curva ROC produzido pelos dois programas é praticamente o mesmo. Da análise dos programas testados para o estudo de um único conjunto de dados, verifica-se que o ROCNPA acrescenta para além da facilidade na introdução dos dados, só igualável a um pacote estatístico como o SPSS, uma análise gráfica mais completa que o AccuROC e uma análise do índice área abaixo da curva ROC tão completa como no ROCKIT. Para além destas características, o ROCNPA fornece um ajuste à curva ROC directamente da análise dos dados.

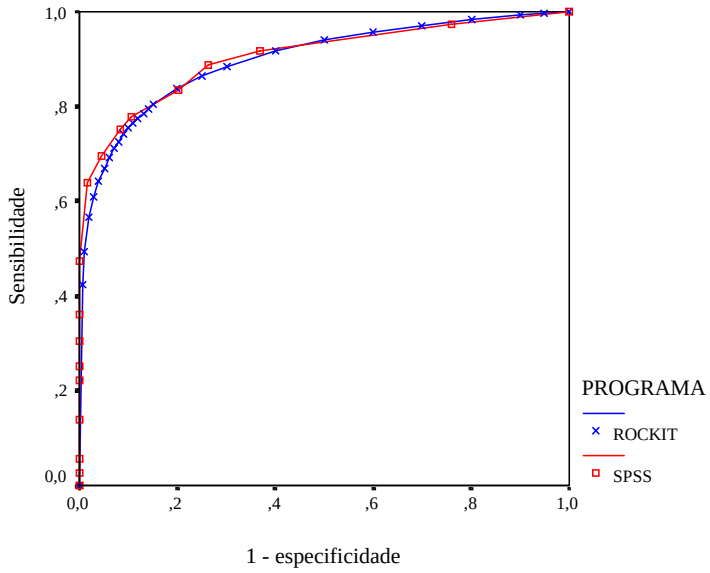


Figura 8.8: Comparação das curvas ROC ajustadas produzidas pelo SPSS e pelo ROCKIT.

8.5.2 Análise de dois ou mais conjuntos de dados correlacionados

Para analisar dois conjuntos de dados correlacionados, no contexto da análise ROC, utilizou-se os dados referentes às variáveis IGN (Idade Gestacional Neonatal) e IGO (Idade Gestacional Obstétrica), já estudadas no capítulo 6. Trata-se de variáveis numa escala ordinal que varia entre 20 e 40 semanas, e um menor valor da escala indica que o recém-nascido irá falecer (teste positivo), enquanto que um maior valor da escala indicará que o recém-nascido irá sobreviver (teste negativo). Dos dados relativos aos 223 bebés, 194 sobreviveram (classificados como sobrevivos), tendo sido registado 29 óbitos (classificados como falecidos). A análise ROC será efectuada utilizando os programas ROCKIT, CLABROC, AccuROC e ROCNPA.

Quando se cria o ficheiro de dados, para este conjunto de variáveis, há que ter em conta que se trata de uma variável cujo menor valor da escala, corresponde ao teste positivo. O resultado da análise no programa ROCKIT encontra-se no anexo C.

No CLABROC, versão para WINDOWS, a análise é igual à efectuada pelo ROCKIT, o ficheiro de dados é que é diferente. Neste programa, apenas é permitido a análise de dois conjuntos de dados correlacionados.

Em termos de análise de resultados, a tabela 8.2 apresenta o resumo dos valores obtidos em cada um dos programas testados, para o índice área abaixo da curva ROC e respectivos erros padrão, considerando a variável CRIB.

Tabela 8.2: Resumo dos valores obtidos para o índice área abaixo da curva ROC

CLABROC	**ROCKIT**	**AccuROC**	**ROCNPA**
A_z(IGN) = 0.818	A_z(IGN) = 0.819		A_z(IGN) = 0.819
SE(A_z) = 0.048	SE(A_z) = 0.052		SE(A_z) = 0.052
A_z(IGO) = 0.811	A_z(IGO) = 0.815		A_z(IGO) = 0.815
SE(A_z) = 0.046	SE(A_z) = 0.048		SE(A_z) = 0.048
	A(IGN) = 0.812	A(IGN) = 0.812	A(IGN) = 0.812
	SE(A) = 0.050	SE(A) = 0.053	SE(A) = 0.050
	A(IGO) = 0.803	A(IGO) = 0.803	A(IGO) = 0.803
	SE(A) = 0.051	SE(A) = 0.048	SE(A) = 0.051

Na tabela 8.3 encontra-se o resumo dos resultados obtidos da comparação das duas escalas, em termos do teste ao índice área abaixo da curva ROC, tendo em conta que se trata de dados correlacionados. O AccuROC efectua a comparação utilizando o método dos contrastes e a estatística do quiquadrado definida em Delong e DeLong [22].

Tabela 8.3: Resumo dos testes de comparação para IGN e IGO.

CLABROC	**ROCKIT**	**AccuROC**	**ROCNPA**
$z = 0.2915$	$z = 0.1538$	$Q = 0.0557$	$-$
$p = 0.7707$	$p = 0.8778$	$p = 0.8134$	$p = 0.734$

A análise destes valores permite concluir que os resultados das comparações múltiplas conduzem ao mesmo tipo de decisão independentemente do teste utilizado.

Quando se pretende comparar mais do que três conjuntos de dados correlacionados, o ROCKIT encrava e o AccuROC não o permite. Assim o ROCNPA apresenta a vantagem de poder efectuar os cálculos de uma forma simples e rápida, apresentando uma folha de resultados que pode ser guardada num formato HTML. O formato HTML é um formato universalmente difundido, regulamentado (www.w3c.org) e aberto, ou seja, não só existem diversos programas capazes de o ler e mostrar como também é possível examinar o seu conteúdo em qualquer simples editor de texto.

8.5.3 Análise de dois ou mais conjuntos de dados independentes

Para ilustrar a análise de dois ou mais conjuntos de dados independentes através de curvas ROC, utilizou-se o conjunto de dados referentes à comparação de desempenho em termos de cuidados prestados para os quatro hospitais estudados do capítulo 6. Trata-se de um conjunto de quatro variáveis independentes, pelo que à partida reduz as opções em termos de programas

disponíveis. A análise foi efectuada no ROCKIT e no ROCNPA. Verificou-se que apenas o ROCNPA apresentou resultados, pois o ROCKIT tornou a apresentar problemas.

8.6 Discussão e conclusão

Como pode ser verificado através dos resultados obtidos, o ROCNPA apresenta praticamente os mesmos resultados em termos de valores de área abaixo da curva ROC e respectivos erros padrão. Tal facto é devido à utilização das mesmas metodologias empregues nos outros programas. A abordagem não paramétrica utilizada, a aproximação à estatística de Wilcoxon-Mann-Whitney, é também a utilizada no ROCKIT e no AccuROC e, por outro lado, a abordagem paramétrica, aproximação à Normal, apesar de utilizar um método de estimação de parâmetros diferente, conduz a resultados semelhantes.

Quanto à capacidade de cálculo dos programas testados, verificou-se que o ROCNPA apresenta maior versatilidade em termos do número de variáveis, quer para o caso de amostras correlacionadas, quer de amostras independentes.

Os resultados obtidos no ROCNPA podem ser facilmente transportados para qualquer processador de texto, dado que os gráficos podem ser guardados num formato de imagem do tipo *.gif* ou então directamente colocados no texto por um sistema $copiar - colar$. A folha de resultados analíticos apresenta um formato livre, HTML, que como referido anteriormente, permite o seu tratamento em qualquer tipo de processador de texto.

Saliente-se ainda, que devido aos requisitos da linguagem de programação utilizada, o JAVA, o ROCNPA confere maior facilidade quer em termos de

introdução de variáveis, quer na análise de resultados.

Conclusão

O estudo que acabou de ser descrito incidiu fundamentalmente sobre três questões básicas:

- procura de uma expressão analítica que traduza a curva ROC;
- tratamento de algumas aplicações através da metodologia ROC;
- desenvolvimento de um programa para a análise ROC.

No que diz respeito à primeira questão, o desenvolvimento foi efectuado no capítulo 5, secção 5.1. A análise apresentada, foi feita através de estudos de simulação, partindo do pressuposto que as duas distribuições, as dos casos *normais* e a dos *anormais*, tinham a mesma forma funcional. Concluiu-se que a relação não linear encontrada na representação no plano unitário, por via da transformação em escalas de probabilidade normal, só produz uma recta quando a distribuição subjacente aos dados segue uma distribuição Normal. Em todos os outros casos estudados, tal transformação produziu relações não lineares.

Este estudo permitiu ainda concluir que no caso da Normalidade, a representação no plano *binormal*, permite, retirar os parâmetros de interesse relativos às duas distribuições através da ordenada na origem e do declive da recta no plano *binormal*.

O estudo efectuado sobre a aplicação da análise através da curva ROC a casos reais, como o da avaliação do risco de morte em recém-nascidos de muito baixo peso, permitiu avaliar de entre cinco tipos diferentes de índices de gravidade clínica qual o mais indicado para determinação do risco de morte para este grupo de recém-nascidos. Nesta avaliação, as variáveis em estudo são correlacionadas, tendo sido determinado o valor do índice área abaixo da curva ROC e respectivos erros padrão.

A avaliação do desempenho dos cinco índices estudados não foi conclusiva quanto ao que poderá apresentar melhor performance, dado que o resultado dos testes de comparação múltipla para o índice área abaixo da curva ROC, não permitiu detectar diferenças estatisticamente significativas. No entanto, a avaliar pela complexidade das escalas (em termos do número de variáveis a recolher e tempo de recolha), em relação ao *CRIB* e, pelo facto deste índice apresentar um maior valor de área abaixo da curva ROC e menor erro padrão, sugere-se que o mesmo poderá ser considerado o melhor índice indicativo do risco de mortalidade neonatal.

Numa outra aplicação, para amostras correlacionadas, estudou-se a *Idade Gestacional* como medida de prognóstico. Nesta situação, o objectivo da análise de diagnóstico é, não só, determinar a influência da idade gestacional como factor de prognóstico no parto (bebé falecido ou sobrevivo), mas também comparar as duas medidas de idade gestacional, *IGO* e *IGN,* avaliando se alguma das escalas é superior.

A comparação das áreas abaixo das curvas ROC, para o estudo efectuado, nada permitiu concluir quanto à melhor medida de avaliação do risco de morte para os bebés nos dois casos analisados, os recém-nascidos de muito baixo peso e o conjunto de todos os recém-nascidos, tal como é confirmado pelos testes à diferença das áreas. No entanto, pelos valores de área abaixo

da curva ROC, verifica-se que a idade gestacional pode ser considerada como um factor de prognóstico importante para a sobrevivência dos bebés, sendo assim um indicador importante na tomada de decisão sobre a indução de um parto.

Com base nos resultados obtidos no primeiro estudo - que permitiu identificar o *CRIB* como sendo o melhor indicador do risco de mortalidade neonatal para recém-nascidos de muito baixo peso, quer devido à menor complexidade de recolha de variáveis quer pelo valor da área abaixo da curva ROC - foi sugerido que se fizesse uma comparação dos cuidados oferecidos por unidades de cuidados intensivos neonatais (UCIN) de vários hospitais, usando o *CRIB* como medida de risco neonatal inicial. Assim, foram avaliadas quatro UCIN de hospitais portugueses, utilizando a metodologia das curvas ROC.

A análise estatística através de comparações múltiplas, para os quatro hospitais, permitiu averiguar que não existiam diferenças significativas em termos de desempenho de cuidados intensivos neonatais, entre a UCIN do hospital H_1 e as dos hospitais H_2 e H_3, sendo significativa apenas a diferença entre a UCIN do hospital H_1 e a do hospital H_4.

Foi ainda realizada uma outra análise referente à associação entre o índice *CRIB* com o aparecimento de três sequelas nos recém-nascidos sobrevivos. A análise das curvas ROC para as três sequelas, segundo o *CRIB* para os indivíduos sobrevivos, verificou-se que a sequela ROP apresenta maior valor de área ($A = 0.82$) e menor erro padrão ($SE(A) = 0.05$), o que demonstra a utilidade deste índice como indicador do risco de aparecimento destas sequelas.

Por fim, devido às dificuldades de cálculo surgidas no decorrer do capítulo 6, e ainda às limitações impostas pelos programas existentes para análise através de curvas ROC, desenvolveu-se um novo programa, o ROCNPA.

O ROCNPA permitiu colmatar algumas lacunas existentes no campo da análise através de curvas ROC, nomeadamente:

- realização dos cálculos de uma forma simples e rápida;

- apresentação de uma grande componente gráfica - curva ROC empírica no plano unitário, curva ROC no plano *binormal* e curva ROC ajustada no plano unitário. Permite ainda, a sobreposição de mais do que três curvas ROC empíricas no mesmo plano unitário;

- apresentação de uma folha de resultados - valor da área abaixo da curva ROC calculada através da regra do trapézio, da aproximação à estatística de Wilcoxon-Mann-Whitney e da aproximação *binormal,* apresentando também o valor do erro padrão respectivo.

Para a comparação entre vários testes, para o caso de amostras independentes ou correlacionadas, são apresentados os valores de prova das comparações múltiplas dois a dois. Para o caso de amostras correlacionadas apresenta ainda, as matrizes de covariância e correlação calculadas segundo o procedimento de DeLong e DeLong.

Há ainda que salientar que o ROCNPA é o único programa para análise através de curvas ROC que pode ser utilizado num outro tipo de plataforma que não o WINDOWS ou DOS, o que o poderá tornar mais atractivo, especialmente para utilizadores de outros sistemas operativos, como por exemplo o LINUX, UNIX e Macintosh.

É sabido, que qualquer trabalho de investigação, nomeadamente aquele que envolve um projecto de doutoramento, deve ter um fim, sob pena de se

arrastar indefinidamente e de não constituir mais do que um motivo de satisfação intelectual para quem nele está envolvido. Por conseguinte, a melhor retribuição que um investigador pode ter quando apresenta um trabalho que passou anos a desenvolver, é a de sentir que ele não se esgota em si próprio e abre portas para nova reflexão. Assim, um investigador ao pôr termo aos seus estudos, deverá ter a consciência de que o domínio em que trabalhou terá muito para explorar. Nesta mensagem procurarei apresentar algumas propostas que poderão servir de objecto de trabalhos futuros de investigação dentro do domínio da análise através de curvas ROC.

Em primeiro lugar, a abordagem não paramétrica à análise ROC parece recolher a preferência da larga maioria dos trabalhos mais recentes. Contudo, existem algumas questões que necessitam de ser aprofundadas e que constituem seguramente propostas de investigação futura, nomeadamente o cálculo dos erros padrão associados às áreas para amostras independentes, a comparação entre os métodos propostos por Hanley e McNeil [38] e DeLong e DeLong [22] para o cálculo do coeficiente de correlação e o estudo dos métodos de estimação dos parâmetros da recta no plano *binormal*.

Tal como foi apresentado nos estudos de simulação realizados (ver secção 5.1.1) é possível gerar curvas que cruzam a diagonal principal, por vezes designadas por curvas ROC impróprias. Se bem que, para dados normais, o cruzamento seja indicador de variâncias diferentes para os casos normais e anormais, é fundamental estudar as implicações, nomeadamente sobre a utilidade do teste em causa.

Por outro lado, na comparação entre testes alternativos é usual encontrar curvas ROC empíricas que se cruzam, questão que continua em aberto. Apesar de ser sempre possível definir qual o teste preferível para diferentes gamas de sensibilidade e especificidade, a definição de um índice de avaliação

global continua por fazer.

No domínio do estudo da forma da curva ROC e das distribuições associadas aos dados, foi apenas tratada a situação de distribuições com a mesma forma funcional. No entanto, poderá constituir uma linha de investigação futura o estudo de mistura de distribuições.

A análise ROC pode ser também percepcionada como um processo de ajuda na tomada de decisão, nomeadamente na escolha entre testes alternativos. Contudo, uma área de grande potencial, e só levemente aflorada neste trabalho, é a comparação entre Unidades de Cuidados Intensivos, contribuindo para a avaliação da sua performance. Potencialmente, a análise ROC também pode contribuir para comparar/treinar a capacidade de diagnóstico de médicos, em particular, nos diagnósticos que impliquem a avaliação de imagens.

Termino este trabalho com um pensamento de *Julien Huxley*, que revela um pouco do que é o espírito científico:

"Uma das coisas para que a ciência serve -
é para nos dar ideia da nossa ignorância".

Apêndice A

Determinação das EMV dos parâmetros na Teoria de Detecção de Sinal, para dados agrupados em classes

O procedimento para obtenção de estimativas de máxima verosimilhança para os parâmetros na Teoria de Detecção de Sinal, para dados agrupados em classes, foi desenvolvido em 1969 por Dorfman e Alf em [27].

Considere-se o modelo para dados agrupados em classes como descrito por Dorfman e Alf [27], em que os acontecimentos experimentais são constituídos por duas classes de estímulos, S_1 e S_2, e por um conjuntode respostas R_j $(j = 1, ..., n' + 1)$.

Axioma A.1 *Em cada experiência, a introdução de um S_i conduz a um acontecimento x situado num espaço unidimencional contínuo.*

Axioma A.2 *Para um conjunto infinito de experiências, a introdução de*

um S_i está associada a uma distribuição Normal dos acontecimentos x com média μ_i e variância σ_i^2.

Axioma A.3 *Existe um conjunto de valores de corte Z_k $(k = 1, ..., n')$, tal que:*

(i) $x < Z_1$ a resposta é R_1,

(ii) $x > Z_{n'}$ a resposta é $R_{n'+1}$,

(iii) $Z_k < x < Z_{k+1}$, a resposta é R_{k+1} $(j = k + 1)$ para todo $k < n'$.

Axioma A.4 *As experiências são consideradas mutuamente independentes. Destes axiomas resulta que*

$$P(R_j \mid S_1) = F(Z_{k=j}) - F(Z_{k=j-1}) \tag{A.1}$$

onde $Z_k = (x_k - \mu_1)/\sigma_1$, F é a função de distribuição acumulada da Normal, $F(Z_0) = 0$, e $F(Z_{n'+1}) = 1$.

$$P(R_j \mid S_2) = F(bZ_{k=j} - a) - F(bZ_{k=j-1} - a), \tag{A.2}$$

onde $b = \sigma_1/\sigma_2$, e $a = (\mu_2 - \mu_1)/\sigma_2$.

Maximiza-se a função de verosimilhança em ordem aos parâmetros a, b e todos os Z_k's, efectuando a diferenciação do logarítmo da função de verosimilhança em ordem a a, b e todos os Z_k, igualando estas expressões a zero, e resolvendo este conjunto de equações. Para os dados agrupados em classes, o logarítmo da função de verosimilhança é dado por:

$$\log L = \sum_{i=1}^{2} \sum_{j=1}^{n'+1} r_{ij} \log P_{ij}, \tag{A.3}$$

onde r_{ij} é o número de R_j's ao estímulo i, e P_{ij} é a probabilidade de R_j dado S_i.

Diferenciando a equação (A.3) em ordem ao parâmetro a, depois de substituir as equações (A.1) e (A.2) em (A.3), obtêm-se:

$$\frac{\partial \log L}{\partial a} = -n_2 \sum_{j=1}^{n'} f(bZ_j - a) \left[\frac{r_{2,j}/n_2}{F_{2,j} - F_{2,j-1}} - \frac{r_{2,j+1}/n_2}{F_{2,j+1} - F_{2,j}} \right], \qquad \text{(A.4)}$$

onde $F_{2,j} = F(bZ_j - a)$, $F_{1,j} = F(Z_j)$ e n_i é o número de s_i's. Diferenciando a equação (A.3) em ordem ao parâmetro b, resulta:

$$\frac{\partial \log L}{\partial b} = n_2 \sum_{j=1}^{n'} f(bZ_j - a)\,(Z_j) \left[\frac{r_{2,j}/n_2}{F_{2,j} - F_{2,j-1}} - \frac{r_{2,j+1}/n_2}{F_{2,j+1} - F_{2,j}} \right]. \qquad \text{(A.5)}$$

Diferenciando equação (A.3) em ordem a Z_k, vem:

$$\begin{aligned} \frac{\partial \log L}{\partial Z_k} &= n_2 f(bZ_j - a)(b) \left[\frac{r_{2,j}/n_2}{F_{2,j} - F_{2,j-1}} - \frac{r_{2,j+1}/n_2}{F_{2,j+1} - F_{2,j}} \right] \\ &\quad + n_1 f(Z_j) \left[\frac{r_{1,j}/n_1}{F_{1,j} - F_{1,j-1}} - \frac{r_{1,j+1}/n_1}{F_{1,j+1} - F_{1,j}} \right]. \end{aligned} \qquad \text{(A.6)}$$

Igualando estas derivadas parciais a zero, obtém-se um conjunto de equações não lineares, cuja solução poderá ser obtida por uma adaptação do método de Newton-Raphson, por vezes designado por *método de scoring* [27].

Especificamente, dado um vector de estimativas consistente, mas insuficiente, um vector de estimativas melhorado é obtido a partir de:

$$\mathbf{S}_1 = \mathbf{S}_0 + \mathbf{A}^{-1}\mathbf{r}, \qquad \text{(A.7)}$$

onde $\mathbf{S}_0$ é o vector de estimativas, $\mathbf{S}_1$ é o vector de estimativas melhorado, $\mathbf{r}$ é o vector das primeiras derivadas parciais com as estimativas iniciais substituidas pelas desconhecidas, e $\mathbf{A}^{-1}$ é a matriz inversa de $\{-E\left(\partial^2 \log L/\partial\theta_1\partial\theta_2\right)\}$. Depois do processo iterativo estar completo, $\mathbf{A}^{-1}$ é a matriz de variância-covariância [27].

As segundas derivadas, são dadas pelas seguintes expressões:

$$E\left[\frac{\partial^2 \log L}{\partial a^2}\right] = -n_2 \sum_{j=1}^{n'} f_{2,j} \left[\frac{f_{2,j} - f_{2,j-1}}{F_{2,j} - F_{2,j-1}} - \frac{f_{2,j+1} - f_{2,j}}{F_{2,j+1} - F_{2,j}}\right],$$

$$E\left[\frac{\partial^2 \log L}{\partial b^2}\right] = -n_2 \sum_{j=1}^{n'} f_{2,j} Z_j \left[\frac{f_{2,j} Z_j - f_{2,j-1} Z_{j-1}}{F_{2,j} - F_{2,j-1}} - \frac{f_{2,j+1} Z_{j+1} - f_{2,j} Z_j}{F_{2,j+1} - F_{2,j}}\right].$$

$$\begin{aligned} E\left[\frac{\partial^2 \log L}{\partial Z_k^2}\right] = & -n_2 f_{2,j} b^2 \left[\frac{f_{2,j}}{F_{2,j} - F_{2,j-1}} - \frac{f_{2,j}}{F_{2,j+1} - F_{2,j}}\right] \\ & - n_1 f_{1,j} \left[\frac{f_{1,j}}{F_{1,j} - F_{1,j-1}} - \frac{f_{1,j}}{F_{1,j+1} - F_{1,j}}\right]. \end{aligned}$$

$$E\left[\frac{\partial^2 \log L}{\partial a \ \partial b}\right] = n_2 \sum_{j=1}^{n'} f_{2,j} \ Z_j \left[\frac{f_{2,j} - f_{2,j-1}}{F_{2,j} - F_{2,j-1}} - \frac{f_{2,j+1} - f_{2,j}}{F_{2,j+1} - F_{2,j}}\right].$$

$$E\left[\frac{\partial^2 \log L}{\partial a \ \partial Z_{k=j}}\right] = n_2 \ f_{2,j} \ b \left[\frac{f_{2,j} - f_{2,j-1}}{F_{2,j} - F_{2,j-1}} - \frac{f_{2,j+1} - f_{2,j}}{F_{2,j+1} - F_{2,j}}\right].$$

$$E\left[\frac{\partial^2 \log L}{\partial b \ \partial Z_{k=j}}\right] = -n_2 \ f_{2,j} \ b \left[\frac{f_{2,j} - f_{2,j-1}}{F_{2,j} - F_{2,j-1}} - \frac{f_{2,j+1} - f_{2,j}}{F_{2,j+1} - F_{2,j}}\right].$$

$$E\left[\frac{\partial^2 \log L}{\partial Z_{k_m} \ \partial Z_{k_n}}\right] = 0, \qquad m \neq n.$$

Estimativas consistentes, mas insuficientes de Z_k´s podem ser obtidas pela resolução do seguinte conjunto de equações:

$$\left(\sum_{j=1}^{k} \widehat{P}_{1j}\right)^{-1} = Z_k \tag{A.8}$$

onde $\widehat{P}_{ij} = r_{ij}/n_i$, e$\left(\sum \widehat{P}_{ij}\right)^{-1}$ significa a transformação inversa de F (desvios normais padronizados).

Para obter estimativas consistentes de a e b, considera-se o seguinte par de equações:

$$\left(\sum_{j=1}^{k} \widehat{P}_{2j}\right)^{-1} = bZ_k - a. \tag{A.9}$$
$$\left(\sum_{j=1}^{k+1} \widehat{P}_{2j}\right)^{-1} = bZ_{k+1} - a.$$

Substituindo Z_k e Z_{k+1} nestas equações, resolvendo em ordem a a e b, e fazendo a média das soluções para cada par de equações ao longo de todos conjuntos de pontos, obtêm-se as estimativas consistentes de a e b.

Apêndice B

Teste de Wilcoxon-Mann-Whitney

As técnicas não paramétricas apresentam várias vantagens sobre os métodos tradicionais de inferência estatística. Uma vantagem é que não incorporam todas as hipóteses restritivas características dos testes paramétricos. Por outro lado, os testes não paramétricos trabalham com graduações em vez dos valores das observações, o que os torna mais simples e rápidos, em termos de cálculo, para pequenas amostras.

No entanto, o uso das graduações torna as técnicas não paramétricas menos sensíveis às medidas do erro de tipo I do que os testes tradicionais [66].

O teste de Mann-Whitney é um teste não paramétrico para comparação de duas distribuições e foi primeiramente introduzido para o caso em que $n_1 = n_2$ por Wilcoxon (1945). O teste de Wilcoxon foi expandido para o caso de amostras com dimensão diferente por White (1952) e Van der Reyden (1952). Um teste equivalente ao de Wilcoxon foi também desenvolvido independentemente e introduzido por Festinger (1946).

Mann e Whitney (1947) parecem ter sido os primeiros a considerar amostras de diferentes tamanhos e a fornecer tabelas para usar com amostras de pequena dimensão [19].

Dado que o teste é atribuído a vários autores, existe a interrogação de que nome lhe atribuir, pelo que se utilizará a designação de Wilcoxon-Mann-Whitney por terem sido estes os primeiros a desenvolverem este tipo de teste não paramétrico.

B.1 Hipóteses

Quando são efectuadas medições ordinais, o teste *U de Mann-Whitney* pode ser usado para testar se dois grupos independentes foram retirados da mesma população. Este é um dos testes mais potentes para comparação de duas distribuições e é a alternativa mais utilizada em relação ao teste paramétrico, o *teste t*.

Supondo que se têm duas populações, *população A* e *população B*, a hipótese nula é:

H_0 : As duas distribuições de probabilidade A e B são idênticas.

Uma hipótese alternativa direccional, H_1, contra a qual se pode testar H_0 é

H_1 : A distribuição de A é estocasticamente maior do que B.

Deve-se aceitar H_1 se a probabilidade de uma classificação de A ser maior do que uma classificação de B, for superior a $\frac{1}{2}$. Isto é, se a for uma observação da população A, e b uma observação da população B, então H_1 é tal que $P(a > b) > \frac{1}{2}$. Se a evidência suportar H_1, isto implica que o a distribuição

dos valores correspondentes à população A se encontram à direita dos da população B (situação semelhante à ilustrada na figura 2.1 do capítulo 2, em que B corresponderia a S_0 e A a S_1). Ter-se-ía o caso em que a média da distribuição A estaria à direita da de B.

Claro, que também se pode prever a situação contrária, isto é, considerar como hipótese alternativa:

H_1: A distribuição de B é estocasticamente maior do que A.

Neste caso H_1 deverá ser tal que $P(a > b) < \frac{1}{2}$.

Para o teste bilateral, isto é, para a previsão de diferenças que não impliquem direcção, H_1 deverá ser tal que $P(a > b) \neq \frac{1}{2}$.

B.2 Método

Seja n_1 o número de casos no menor dos dois grupos independentes, e n_2 o número de casos no maior. Para aplicar o *teste U*, primeiro combinam-se as observações ou classificações dos dois grupos, e atribuem-se as graduações de uma forma crescente.

De seguida, escolhe-se um grupo, por exemplo, o grupo com n_1 casos. O valor de U (estatística usada no teste) é dado pelo número de vezes que a classificação no grupo com n_2 casos precede a classificação no grupo com n_1 casos na graduação.

Considere-se um exemplo em que existem dois grupos, um designado por "*anormal*", **A**, com três casos, e um outro designado por "*normal*", **N**, com quatro casos. Nesta situação, $n_1 = 3$ e $n_2 = 4$. Suponha-se ainda, que se registam as seguintes observações:

Grupo A	9	11	15	
Grupo N	5	8	10	12

Para determinar U, começa-se por ordenar estas observações por ordem crescente, da seguinte forma

5	8	9	10	11	12	15
N	**N**	**A**	**N**	**A**	**N**	**A**

Agora considere-se o grupo designado por **N**, e conte-se o número de classificações **A** que precede cada classificação no grupo **N**.

Assim, neste exemplo ter-se-ía:

$$U = 0 + 0 + 1 + 2 = 3$$

o número de vezes que a classificação **A** precede a classificação **N** é 3.

A distribuição amostral de U sob H_0 é conhecida, e com este conhecimento pode-se determinar a probabilidade associada à ocorrência sob H_0 de qualquer U como extremo de um valor observado de U.

B.3 Amostras de dimensão reduzida

Quando nem n_1 nem n_2, são maiores do que 8, a tabela J existente em [77], pode ser utilizada para determinar a probabilidade associada à ocorrência sob H_0 de qualquer U como extremo de um valor observado de U.

Para determinar a probabilidade sob H_0 associada aos seus dados, o investigador apenas precisa de conhecer n_1 (dimensão do grupo mais pequeno), n_2 e U. Com esta informação ele pode ler o valor de p da tabela apropriada para os seus valores.

No exemplo considerado anteriormente, para $n_1 = 3$, $n_2 = 4$ e $U = 3$, retira-se da tabela J que $P(U \leq 3) = 0.200$.

Os valores das probabilidades apresentados na tabela J são para o teste unilateral. Para o teste bilateral, o valor de p retirado da tabela deveria ser multiplicado por dois.

Para valores de n_1 e n_2 elevados, o método de contagem para determinar o valor de U poderá ser um pouco fastidioso. Um método alternativo, que dá resultados idênticos, é atribuir graduação 1 ao valor mais pequeno do conjunto $(n_1 + n_2)$ do grupo de classificações e assim sucessivamente. Então,

$$\begin{aligned} U_2 &= W_2 - \frac{n_2\,(n_2+1)}{2} \\ U_2 &= n_1 n_2 + \frac{n_1\,(n_1+1)}{2} - W_1 \end{aligned} \tag{B.1}$$

ou equivalentemente

$$\begin{aligned} U_1 &= W_1 - \frac{n_1\,(n_1+1)}{2} \\ U_1 &= n_1 n_2 + \frac{n_2\,(n_2+1)}{2} - W_2 \end{aligned} \tag{B.2}$$

onde W_1 representa a soma das graduações para o grupo 1 e W_2 representa a soma das graduações para o grupo 2.

Pode-se transformar U_1 em U_2 através da expressão:

$$U_2 = n_1 n_2 - U_1 \tag{B.3}$$

consequentemente

$$P(U_2 \geq U_1) = P(U_2 \leq n_1 n_2 - U_1).$$

Existe um outro tipo de tabela que dá o valor do ponto crítico correspondente a um determinado valor de U, para valores de n_2 entre 9 e 20 (tabela K [77]).

A distribuição de base da tabela K assenta no menor dos valores entre U_1 e U_2, usualmente designado por U.

B.4 Amostras de grande dimensão

Foi demonstrado por Mann e Whitney (1947), que quando n_1 e n_2 aumentavam em dimensão, a distribuição de U rapidamente se aproxima da distribuição Normal, com

$$\mu_U = \frac{n_1 n_2}{2}$$

e desvio padrão

$$\sigma_U = \sqrt{\frac{n_1 n_2 \left(n_1 + n_2 + 1\right)}{12}}$$

Assim, quando $n_2 > 20$, pode-se determinar a significância de um valor observado U, através de

$$z_U = \frac{U - \mu_U}{\sigma_U} \sim N(0,1)$$

$$z_U = \frac{U - \frac{n_1 n_2}{2}}{\sqrt{\frac{n_1 n_2 (n_1+n_2+1)}{12}}} \tag{B.4}$$

Quando se faz a aproximação à Normal para a distribuição de U, não interessa qual das fórmulas se utiliza para o cálculo do valor de U, pois o valor absoluto de z dado pela expressão da equação B.4 será o mesmo independentemente da expressão utilizada no cálculo de U. O sinal de z depende de que valor se utiliza, U ou U', mas o valor não.

B.5 Observações Repetidas

Numa experiência científica é usual se verificarem observações repetidas. Para o teste de Wilcoxon-Mann-Whitney, se as repetições ocorrem no mesmo grupo, o valor de U não é afectado, mas se estas ocorrem em grupos diferentes o valor de U já é afectado. No entanto este efeito é insignificante. Verificando-se a existência de observações repetidas, pode-se utilizar uma correcção que é válida quando se está perante a aproximação à distribuição normal para amostras de grande dimensão.

O efeito dos empates nas graduações é mudar a variabilidade do conjunto das graduações. Assim, a correcção para os empates deverá ser efectuada a nível do desvio padrão da distribuição de U. O desvio padrão corrigido para os empates é dado por [77]

$$\sigma_U = \sqrt{\left(\frac{n_1 n_2}{N(n-1)}\right)\left(\frac{N^3 - N}{12} - \sum T\right)} \tag{B.5}$$

onde $N = n_1 + n_2$

$T = (t^3 - t)/12$ (com t o número de observações repetidas para uma dada graduação)

$\sum T$ é determinado pela soma dos T's para todos os grupos com observações repetidas.

Com a correcção para as observações repetidas o valor de z virá

$$z_U = \frac{U - \frac{n_1 n_2}{2}}{\sqrt{\left(\frac{n_1 n_2}{N(n-1)}\right)\left(\frac{N^3 - N}{12} - \sum T\right)}} \tag{B.6}$$

B.6 Potência de teste

Em termos de potência de teste, poder-se-á comparar este teste não paramétrico com o seu similar em termos paramétricos, o *teste t*.

Se o teste de Mann-Whitney for aplicado a dados que poderiam ser analisados pelo teste paramétrico mais potente, o *teste t*, a sua potência aproxima-se de $95,5\%$ com o aumento de N, e está perto de 95% mesmo para amostras de dimensão moderada. É então, uma excelente alternativa ao *teste t*, não tendo as condições restritivas que se encontram associadas a este teste [77].

Apêndice C

Listagem de resultados obtidos nas comparações de programas para análise ROC

Neste anexo é apresentado um conjunto de listagens resultantes das comparações de programas para análise ROC efectuadas no capítulo 8.5. Pretende-se assim, elucidar o leitor sobre o tipo de comandos que terão de efectuar, para proceder à análise estatística de dados utilizando a curva ROC, ou o tipo de resultados que poderão esperar aquando a utilização de cada um dos programas testados.

C.1 ROCKIT

```
Date - 23-Nov-00
  Time - 15:15:38
```

```
ROCKIT (Windows95 version 0.9 BETA):

Maximum Likelihood Estimation of a Binormal ROC Curve


From CONTINUOUSLY-Distributed Test Results
```

Original input of 133 Actually-NEGATIVE cases

---

6. 1. 5. 0. 3.

1. 6. 1. 1. 3.

0. 4. 8. 3. 3.

1. 4. 9. 0. 1.

2. 4. 1. 1. 2.

1. 1. 1. 1. 1.

4. 1. 1. 0. 1.

1. 4. 0. 0. 1.

1. 2. 1. 1. 0.

1. 1. 2. 1. 2.

8. 5. 2. 0. 0.

4. 4. 2. 1. 3.

3. 1. 2. 9. 1.

0. 0. 0. 0. 0.

1. 1. 1. 0. 5.

0. 8. 0. 0. 1.

1. 6. 1. 3. 6.

1. 1. 1. 0. 1.

1. 1. 1. 0. 1.

1. 0. 3. 0. 2.

8. 0. 1. 1. 1.

4. 0. 1. 6. 0.

0. 1. 4. 2. 2.

1. 1. 0. 2. 2.

1. 1. 0. 4. 4.

4. 2. 0. 0. 0.

4. 1. 0.

Original input of 36 Actually-POSITIVE cases

---

8. 2. 6. 9. 10.

10. 13. 4. 9. 8.

6. 9. 7. 11. 4.

1. 13. 12. 14. 11.

8. 8. 9. 15. 7.

16. 3. 5. 8. 8.

3. 1. 0. 14. 14.

13.

Date - 23-Nov-00

Time - 15:15:39

ROCKIT (Windows95 version 0.9 BETA):

CRIB

Maximum Likelihood Estimation of the Parameters

a Single Binormal ROC Curve

Name of Input File being used: CRIB.dat

Condition 1: CRIB

Total number of actually-negative cases = 133.

Total number of actually-positive cases = 36.

Data collected on a nominally continuous scale.

Larger values of the test result represent stronger evidence that the case is actually-positive (e.g., that the patient is actually abnormal)

Operating Points Corresponding to the Input Data Categorized by the LABROC5 Scheme:

FPF: .000 .000 .015 .045 .045 .083 .105 .203 .263 .368 .759

TPF: .000 .361 .472 .639 .694 .750 .778 .833 .889 .917 .972

FPF: 1.000

TPF: 1.000

Initial Estimates of the Binormal ROC Parameters:

---

a = 1.5668

b = .6882

z(k) = .698 -.332 -.623 -.835 -1.238 -1.371 -1.653 -1.712 -2.268 -2.876

Procedure Converges after 4 Iterations

==========

Final Estimates of the Binormal ROC Parameters

==========

Binormal Parameters and Area Under the Estimated ROC :

a = 1.5539

b = .6766

Area (Az) = .9009

Area (Wilc) = .8994

1: z(k) = -.700 .329 .622 .839 1.238 1.372 1.656 1.749 2.309 2.830

Estimated Standard Errors and Correlation of these Values:

Std. Err. (a) = .2923

Std. Err. (b) = .1534

Corr(a,b) = .6948

Std. Err. (Az) = .0332

Std. Err.(Wilc)= .0355

Symmetric 95% Confidence Intervals

For a : (.9810, 2.1268)

For b : (.3760, .9773)

Asymmetric 95% Confidence Interval

For Az: (.8197, .9515)

Variance-Covariance Matrix:

==========

a b z(1) z(2) z(3) z(4) z(5) z(6) z(7) z(8) z(9) z(10)

a .0854

b .0312 .0235

z(1) .0053 .0019 .0140

z(2) .0065 .0013 .0058 .0121

z(3) .0067 .0007 .0049 .0101 .0132

z(4) .0068 0.0000 .0043 .0089 .0117 .0146

z(5) .0065 -.0022 .0035 .0074 .0098 .0122 .0195

z(6) .0060 -.0034 .0032 .0070 .0093 .0116 .0186 .0220

z(7) .0044 -.0068 .0027 .0063 .0084 .0107 .0173 .0205 .0299

z(8) .0036 -.0082 .0026 .0061 .0082 .0104 .0170 .0202 .0295 .0335

z(9) -.0057 -.0214 .0014 .0050 .0073 .0097 .0167 .0202 .0302 .0343 .0732

z(10) -.0210 -.0389 .0001 .0041 .0067 .0096 .0179 .0220 .0338 .0388 .0840 .1481

Correlation Matrix:

==========

a b z(1) z(2) z(3) z(4) z(5) z(6) z(7) z(8) z(9) z(10)

a 1.0000

b .6948 1.0000

z(1) .1528 .1054 1.0000

z(2) .2031 .0768 .4484 1.0000

z(3) .2010 .0390 .3578 .7982 1.0000

z(4) .1927 -.0006 .3003 .6725 .8431 1.0000

z(5) .1585 -.1035 .2109 .4833 .6092 .7249 1.0000

z(6) .1395 -.1478 .1846 .4299 .5440 .6490 .8983 1.0000

z(7) .0874 -.2548 .1340 .3307 .4244 .5108 .7159 .7992 1.0000

z(8) .0672 -.2921 .1191 .3026 .3908 .4724 .6660 .7446 .9337 1.0000

z(9) -.0716 -.5145 .0449 .1689 .2345 .2961 .4429 .5025 .6443 .6933 1.0000

z(10) .0000 .0000 .0000 .0000 .0000 .0000 .0000 .0000 .0000 .0000 .0000 .0000

Estimated Binormal ROC curve, with Lower and Upper

Bounds of the Asymmetric Point-wise 95% Confidence

Interval for True-Positive Fraction at a Variety

of False-Positive Fractions:

FPF TPF (Lower Bound, Upper Bound)

.005 .4249 (.2274 , .6439)

.010 .4918 (.2977 , .6879)

.020 .5651 (.3810 , .7359)

.030 .6106 (.4349 , .7660)

.040 .6439 (.4751 , .7883)

.050 .6703 (.5070 , .8062)

.060 .6920 (.5334 , .8211)

.070 .7106 (.5559 , .8339)

.080 .7267 (.5754 , .8451)

.090 .7410 (.5927 , .8551)

.100 .7538 (.6081 , .8641)

.110 .7654 (.6220 , .8722)

.120 .7760 (.6347 , .8797)

.130 .7857 (.6463 , .8865)

.140 .7947 (.6570 , .8928)

.150 .8031 (.6670 , .8986)

.200 .8376 (.7080 , .9224)

.250 .8638 (.7395 , .9399)

.300 .8848 (.7649 , .9532)

.400 .9166 (.8050 , .9717)

.500 .9399 (.8367 , .9833)

.600 .9577 (.8637 , .9907)

.700 .9718 (.8881 , .9953)

.800 .9831 (.9119 , .9981)

.900 .9923 (.9377 , .9995)

.950 .9962 (.9539 , .9999)

Estimated Relationship between the Critical Test-Result Value
(which separates 'positive' results form 'negative' results)
and the Corresponding Operating Point on the Fitted Binormal

ROC Curve:

****

Critical Test (FPF , TPF)

Result Value

9.5 (.002, .359)

8.5 (.010, .496)

7.5 (.040, .644)

6.5 (.049, .668)

5.5 (.085, .734)

4.5 (.108, .763)

3.5 (.201, .838)

2.5 (.267, .871)

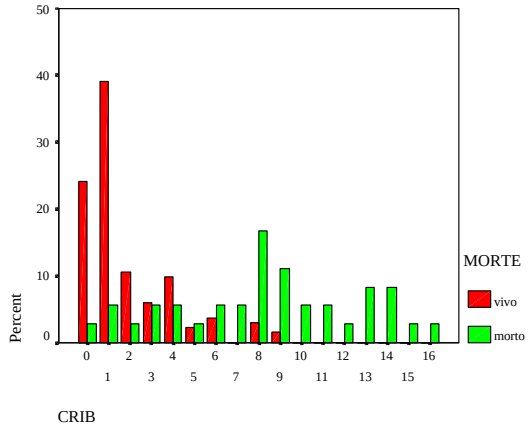
1.5 (.371, .908)

.5 (.758, .979)

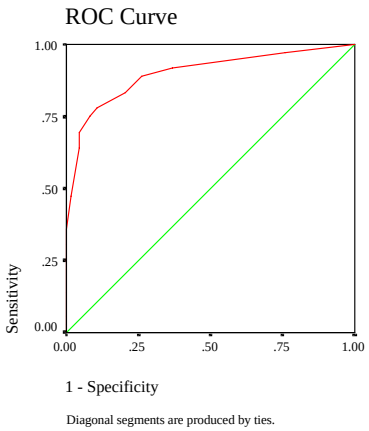
C.2 SPSS

```
CROSSTABS
  /TABLES=crib  BY morte
  /FORMAT= AVALUE TABLES
  /STATISTIC=BTAU
  /CELLS= COUNT
  /BARCHART .


Crosstabs
GRAPH
  /BAR(GROUPED)=PCT BY crib BY morte
```

```
  /MISSING=REPORT.
```




```
ROC
  crib  BY morte (1)
  /PLOT = CURVE(REFERENCE)
  /PRINT = SE COORDINATES
  /CRITERIA = CUTOFF(INCLUDE) TESTPOS(LARGE) DISTRIBUTION(FREE) CI(95)
  /MISSING = EXCLUDE .
Case Processing Summary
MORTE Valid N (listwise)
Positive 36
Negative 133
Larger values of the test result variable(s) indicate stronger
evidence for a positive actual state.
a The positive actual state is morto.


Area Under the Curve
```



CRIB

Area Std. Error Asymptotic Sig. Asymptotic 95% Confidence Interval

Lower Bound Upper Bound

.899 .034 .000 .833 .966

The test result variable(s): CRIB has at least one tie between the positive actual state group and the negative actual state group. Statistics may be biased.

a Under the nonparametric assumption

b Null hypothesis: true area = 0.5

CRIB

Positive if Greater Than or Equal To Sensitivity 1 - Specificity

-1.00 1.000 1.000

.50 .972 .759

1.50 .917 .368

2.50 .889 .263

3.50 .833 .203

4.50 .778 .105

```
5.50 .750 .083
6.50 .694 .045
7.50 .639 .045
8.50 .472 .015
9.50 .361 .000
10.50 .306 .000
11.50 .250 .000
12.50 .222 .000
13.50 .139 .000
14.50 .056 .000
15.50 .028 .000
17.00 .000 .000
```

Bibliografia

[1] A. C. Braga, P. N. Oliveira, A. Gomes. ", A AVALIAÇÃO DO RISCO DE MORTE EM RECÉM-NASCIDOS DE MUITO BAIXO PESO: UMA COMPARAÇÃO BASEADA EM CURVAS ROC." *A Estatística a Decifrar O Mundo*, edited by Luísa Canto e Castro e Dinis Pestana Rita Vasconcelos, Isabel Fraga Alves. Lisboa: Edições Salamandra Lda, 1997.

[2] A. C. Braga, P. N. Oliveira, A. Gomes. ", EVALUATION OF THE RISK OF DEATH FOR VERY LOW BIRTHWEIGHT BABIES A COMPARATION BETWEEN NEONATAL INTENSIVE CARE UNITS: APPLICATION OF ROC CURVES," *Applied Statistical Science IV* (1998).

[3] A. C. Braga, P. N. Oliveira, A. Gomes. "COMPARAÇÃO ENTRE UNIDADES DE CUIDADOS INTENSIVOS NEONATAIS BASEADA NA ANÁLISE ROC." *Estatística: A Diversidade Na Unidade*, edited by Manuela Souto de Miranda e Isabel Pereira. Lisboa: Edições Salamandra Lda, 1998.

[4] A. C. Braga, P. N. Oliveira. "A FORMA DAS CURVAS ROC E A SUA RELAÇÃO COM AS DISTRIBUIÇÕES ASSOCIADAS AOS DADOS." *Afirmar a Estatística. Um Desafio Para O Século XXI*, edited by Ana

Pires e Ferreira Da Cunha Carlos D. Paulino, António Pacheco. Lisboa: Edições SPE, 1999.

[5] Alexander, Greg R., Des Caunes Francois Hulsey Thomas C. Tompkins Mark E and Allen Marilee. "Ethnic Variation in Postnatal Assessments of Gestational Age: A Reappraisal," *Pediatrics and Perinatal Epidemiology*, *6*:423–433 (1992).

[6] Alexander, Greg R., Hulsey Thomas C. Smeriglio Vincent L. Comfort Marilee e Levkoff Abner. "Factors Influencing the Relationship Between a Newborn Assessment of Gestational Maturity and Gestational Age Interval.," *Pediatrics and Perinatal Epidemiology*, *4*:133–146 (1990).

[7] Altman, Douglas G. *Pratical Statistics for Medical Research*. London: Chapman & Hall, 1991.

[8] Bamber, Donald. "The Area Above the Ordinal Dominance Graph and the Area Below the Receiver Operating Characteristic," *Journal of Mathematical Psychology*, *12*:387–415 (1975).

[9] Begg, C. "Advances in Statistical Methodology for Diagnostic Medicine in 1980s," *Statistics in Medicine*, *10*:1887–1895 (1991).

[10] Bland, J. Martin and Douglas G. Altman. "Statistical Methods for Assessing Agreement Between Two Methods of Clinical Measurement," *The Lancet*, 307–310 (1986).

[11] Burdette, J. H., et al. "Alzheimer Disease: Improved Visual Interpretation of PET Images by Using Three-Dimensional Stereotaxic Surface Projections," *Radiology*, *198*:837–843 (1996).

[12] Charles E. Metz, Ben Herman, Jong-Her Shen Helen B. Kronman e Pu-Lan Wang. *CLABROC Program (IBM-PC Version 1.2.1)*. Department of Radiology and Franklin Mclean Memorial Research Institute, University of Chicago, Chicago, Illinois 60637, December 1993.

[13] Charles E. Metz, Benjamin Herman, Jong-Her Shen Helen B. Kronman e Pu-Lan Wang. *ROCKIT 0.9B Beta Version*. Department of Radiology, University of Chicago, Chicago, Illinois 60637, March 1998.

[14] Charles E. Metz, Helen B. Kronman, Pu-Lan Wang e Jong-Her Shen. *INDROC Program (IBM-PC Version)*. Department of Radiology and the Franklin McLean Memorial Research Institute, University of Chicago, Chicago, Illinois 60637, June 1989.

[15] Charles E. Metz, Helen B. Kronman, Pu-Lan Wang e Jong-Her Shen. *ROCPWRPC Program for the IBM PC*. Department of Radiology and the Franklin McLean Memorial Research Institute, University of Chicago, Chicago, Illinois 60637, June 1989.

[16] Charles E. Metz, Helen B. Kronman, Pu-Lan Wang Jong-Her Shen e Ben Herman. *CORROC2 Program (IBM-PC Version 1.2.1)*. Department of Radiology and the Franklin McLean Memorial Research Institute, University of Chicago, Chicago, Illinois 60637, December 1993.

[17] Chen, W. J., et al. "Diagnostic Accuracy of the Child Behavior Checklist Scales for Attention-Deficit Hyperactivity Disorder: A Receiver Operating Characteristic Analysis," *Journal of Consulting and Clinical Psychology*, *62*(5):1017–1025 (1994).

[18] Colliver, J. A., et al. "Screening Test Length For Sequencial Testing with a Standardized-Patient Examination: A Receiver Operating Characteristic (ROC) Analysis," *Academic Medicine*, *67*(9):592–595 (1992).

[19] Conover, W. J. *Pratical Nonparametric Statistics* (2nd ed Edition). New York: John Wiley & Sons, 1971.

[20] Constantine, Norman A., Kraemer Helena C. Kendall-Tackett Katheleen A.-Bennett Forrest C. Tyson Jon E. and Ruth T. Gross. "Use of Physical and Neurologic Observations in Assessment of Gestational Age in Low Birth Weight Infants.," *The Journal of Pediatrics*, *110*(6):921–928 (1987).

[21] Courcy-Wheeler, R. H. B., et al. "Use of the CRIB (Clinical Risk Index for Babies) Score in Prediction of Neonatal Mortality and Morbidity," *Archives of Disease in Childhood*, *73*:F32–F36 (1995).

[22] DeLong, E. R., DeLong D. M. and D. L Clarke-Pearson. "Comparing the Areas Under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach," *Biometrics*, *44*:837–845 (1988).

[23] Diamond, G. A. "Reverend Bayes Silent Majority.An Alternative Factor Affecting Sensitivity and Specificity of Exercise Electrocardiography," *The American Journal of Cardiology*, *57*:1175–1180 (1986).

[24] Dombrowski, Mitchell P., Wolfe Honor M. Brans-Yves W. Saleh-Abdel Aziz A. and Robert J. Sokol. "Neonatal Morphometry. Relation to Obstetric, Pediatric, and Menstrual Estimates of Gestational Age.," *AJDC*, *146*:852–856 (1992).

[25] Donald Dorfman, Kevin Berbaum, Charles E. Metz-Ben Herman e Hatern Abu-Dagga. *LABMRMC 1.0 Beta Version*. University of Chicago, Chicago, Illinois 60637, April 1997.

[26] Dorfman, D. D., Beavers L. L. and C. Saslow. "Estimation of Signal Detection Theory Parameters from Rating-Method Data: A Comparison of the Method of Scoring and Direct Search," *Bull. Psychon. Soc.*, *1*(3):207–208 (1973).

[27] Dorfman, Donald D. and Edward Jr. Alf. "Maximum - Likelihood Estimation of Parameters of Signal - Detection of Confidence Intervals - Ratng-Method Data.," *Journal of Mathematical Psychology*, *6*:487 – 496 (1969).

[28] Egan, James P. *Signal Detection Theory and ROC Analysis*. New York: Academic Press, 1975.

[29] Eskicioglu, Ahmet M. "Application of Multidimensional Quality Measures to Reconstructed Medical Images," *Optical Engineering*, *35*(3):778–785 (1996).

[30] Gagliardi, Luigi, Brambilla C. Bruno Raffaella Martinelli-S. and M. Console. "Biased Assessment of Gestational Age at Birth When Obstetric Gestation is Known.," *Archives of Disease in Childhood*, *68*:32–34 (1993).

[31] Gatsonis, C. and B. J. McNeil. "Collaborative Evaluations of Diagnostic Tests: Experience of the Radiology Diagnostic Oncology Group," *Radiology*, *175*:571–575 (1990).

[32] Goddard, M. J. and I. Hinberg. "Receiver Operator Characteristic (ROC) Curves and Non_Normal Data: An Empirical Study,," *Statistics in Medicine*, *9*:325–337 (1990).

[33] Green, D. M. and J. A. Swets. *Signal Detection Theory and Psychophysics*. New York: Robert E. Krieger Publishing Company, 1973.

[34] H. Ramalho, C. Braga, P. Oliveira A. Alegria. "CRIB: PREDICTIVE ACCURACY AND MORBIDITY," *RELAN (Revista Latinoamericana de Neonatologia)*, *1*(2):111–116 (1999).

[35] Halpern, E.J., et al. "Comparison of Receiver Operating Characteristic Curves on the Basis of Optimal Operating Points," *Statistics for Radiologists*, *3*(3):245–253 (1996).

[36] Hanley, J. A. "The Robustness of "Binormal" Assumptions Used in Fitting ROC Curves," *Medical Decision Making*, *8*:197–203 (1988).

[37] Hanley, J. A. and B. J. McNeil. "The Meaning and Use of the Area under a Receiver Operating Characteristic (ROC) Curve," *Radiology*, *143*:29–36 (1982).

[38] Hanley, J. A. and B. J. McNeil. "A Method of Comparing the Areas under Receiver Operating Characteristic Curves," *Radiology*, *148*:839–843 (1983).

[39] Henkelman, R. M., et al. "Receiver Operator Characteristic (ROC) Analysis Without Truth," *Medical Decision Making*, *10*:24–29 (1990).

[40] Hill, C. C., Rowland D. Y. "Performing ROC Analysis Using S-Plus II," (1998).

[41] Hlatky, M. A., et al. "Rethinking Sensitivity and Specificity," *The American Journal of Cardiology*, *59*:1195–1198 (1987).

[42] Holmes, J. H. "Discovering Risk of Disease with Learning Classifier System," (1997).

[43] Hosmer, Jr, D. W. and S. Lemeshow. *Applied Logistic Regression*. New York: John Wiley & Sons, 1989.

[44] Hsieh, F. and B. W. Turnbull. "Nonparametric and Semiparametric Estimation of the Receiver Operating Characteristic Curve," *The Annals of Statistics*, *24*(1):25–40 (1996).

[45] Hsieh, Fu-Shing. *Performance of Diagnostic Test in a Nonparametric Setting*. PhD dissertation, Cornell University, January 1991.

[46] Ishwaran, H. and C. Gatsonis. "A General Class of Hierarchical Ordinal Regression Models With Applications to Correlated ROC Analysis," (1998?).

[47] Iverson, G. J. and Ching-Fan Sheu. "Characterizing Random Variables in the Context of Signal Theory," *Mathematical Social Sciences*, *23*:151–174 (1992).

[48] Jiang, Y., et al. "Malignant and Benign Clustered Microcalcifications: Automated Feature Analysis and Classification," *Radiology*, *198*:671–678 (1996).

[49] Johnson, N.L., Kotz S. e Balakrishnan N. *Continuous Univariate Distributions* -. New York: John Wiley, 1994.

[50] Kendall, Sir Maurice and Alan Stuart. *The Advanced Theory of Statistics* (4th Edition), *2*.

[51] Kraemer, H. C. "Assessment of 2×2 Associations: Generalization of Signal-Detection Methodology," *The American Statistician*, *42*(1):37–49 (1988).

[52] Martins, F. Mário. *Programação Orientada Aos Objectos Em JAVA*. LIDEL.

[53] McKenzie, D. P. and D. M. Clarke. "Cutoff: A Fortran Program for Establishing Threshols for Screening Indices," *Educational and Psychological Measurement*, *52*:891–893 (1992).

[54] McMillan, S. A., et al. "Evaluation of Formulae for CSF IgG Synthesis Using Data Obtained from Two Methods: Importance of Receiver Operator Characteristic Curve Analysis," *Journal of Clinical Pathology*, *49*:24–28 (1996).

[55] McNeil, B. J., Hanley J. A. Funkenstein H. H. and J. Wallman. "Paired Receiver Operating Characteristic Curves and the Effect of History on Radiographic Interpretation," *Radiology*, *149*:75–77 (1983).

[56] Metz, C. E. "Basic Principles of ROC Analysis," *Seminars in Nuclear Medicine*, *VIII*(4):283–298 (1978).

[57] Metz, C. E. "ROC Methodology in Radiologic Imaging," *Investigative Radiology*, *21*:720–733 (1986).

[58] Metz, C. E. "Statistical Analysis of ROC Data in Evaluating Diagnostic Performance." *Multiple Regression Analysis: Applications in the Health Sciences*, number 13, edited by Donald E. Herbert and Raymond H. Myers. 365–384. American Institute of Physics, 1986.

[59] Metz, C. E. "FORTRAN Programs ROCFIT, CORROC AND ROCPWR." Disponível na Internet, cedido pelo prof. C. Metz, Department of Radiology, University of Chicago, Chicago, IL., 1998.

[60] Metz, C. E., et al. "A New Approach for Testing the Significance of Differences Between ROC Curves Measured from Correlated Data." *Information Processing in Medical Imaging, Proceedings of the 8th Conference*, edited by F. Deconinck. 432–445. Boston: Martinus Nijhoff Publishers, 1983.

[61] Mossman, Douglas. "Assessing Predictions of Violence: Being Accuracy," *Journal of Consulting and Clinical Psychology*, *62*(4):783–792 (1994).

[62] Murtaugh, Paul A. "ROC Curves with Multiple Marker Measurements," *Biometrics*, *51*:1514–1522 (1995).

[63] Murteira, Bento José F. *Probabilidades e Estatística* (2 Edition). Lisboa: McGraw-Hill, 1990.

[64] Network, The International Neonatal. "The CRIB (Clinical Risk Index for Babies) Score:A Tool for Assessing Initial Neonatal Risk and Comparing Perfomance of Neonatal Intensive Care Units," *The Lancet*, *342*:193–198 (1993).

[65] Ott, William J. "Accurate Gestacional Dating: Revisited," *American Journal of Perinatology*, *11*(6):404–408 (1994).

[66] Pagano, Marcello. *Principles of Biostatistics*. Belmont: Duxbury Press, 1993.

[67] Parker, J., et al. "Classification of Ductal Carcinoma in Situ by Image Analysis of Classifications from Digital Mammograms," *The British Journal of Radiology*, *68*:150–159 (1995).

[68] Philbrick, J. T., et al. "Methologic Problems of Exercise Testing for Coronary Artery Disease: Groups, Analysis and Bias," *The American Journal of Cardiology*, *46*:807–812 (1980).

[69] Pollack, I. and R. Hsieh. "Sampling Variability of the Area Under ROC-Curve and of d_e," *Psycological Bulletin*, *71*(3):161–173 (1969).

[70] Pollack, Murray M., Koch Matthew A. Bartel Doris A. Rapoport Irina Dhanireddy-R. El-Mohandes Ayman A. E. Harkavy K. Subramanian K. N. S. and District of Columbia Neonatal Network. "A Comparison of Neonatal Mortality Risk Prediction Models in Very Low Birth Weight Infants," *Pediatrics*, *105*(5):1051–1057 (2000).

[71] Ratcliff, Roger, et al. "Testing Global Memory Models Using ROC Curves," *Psycological Review*, *99*(3):518–535 (1992).

[72] Rautonen, J., et al. "CRIB and SNAP: Assessing the Risk of Death for Preterm Neonates,," *Clinical Pratice, The Lancet*, *343*:1272–1273 (1994).

[73] Ribeiro, M. G., Pinto R. Oliveira P. and M. C. Sá Miranda. "Identification of GM2 - Gangliosidosis B1 Variant Carriers," *J. Inher. Metab. Dis.*, *16*:1003–1011 (1993).

[74] Rifkin, M. D., et al. "Comparison of Magnetic Resonance Imaging and Ultrasonography in Staginc Early Prostate Cancer," *The New England Journal of Medicine*, *323*(10):621–626 (1990).

[75] Rockette, H. E., Obuchowski N. A. and D. Gur. "Nonparametric Estimation of Degenerate ROC Data Ses Used for Comparison of Imaging Systems," *Statistics in Radiology*, *25*(7):835–837 (1990).

[76] Sanders, Marilyn, Allen Marilee Alexander Greg R. Yankowitz Jerome Graeber Janet Johnson-Timothy R. B. and Repka Michael X. "Gestational Age Assessment in Preterm Neonates Weihing Less Than 1500 Grams.," *Pediatrics*, *88*(3):542–546 (1991).

[77] Siegel, Sidney. *Nonparametric Statistics for the Behavioral Sciences*. McGraw-Hill Series in Psychology, 1956.

[78] Swets, J. A. "Form of Empirical ROCs in Discrimination and Diagnostic Tasks," *Psycological Bulletin*, *99*(2):181–198 (1986).

[79] Swets, J. A. "Measuring the Accuracy of Diagnostic Systems," *Science*, *240*:1285–1293 (1988).

[80] Swets, J. A. *Signal Detection Theory and ROC Analysis in Psychology and Diagnostics: Collected Papers*. New Jersey: LEA, 1996.

[81] Swets, J. A. and R. M. Pickett. *Evaluation of Diagnostic Systems: Methods from Signal Detection Theory*. London: Academic Press, 1982.

[82] Swets, J. A., et al. "Assessement of Diagnostic Technologies," *Science*, *205*:753–759 (1979).

[83] Tavel, M. E., et al. "Sensitivity and Specificity of Tests: Can the "Silent Majority" Speak," *The American Journal of Cardiology*, *60*:1167–1169 (1987).

[84] Valenstein, P. N. "Evaluating Diagnostic Tests with Imperfect Standards," *American Journal of Clinical Pathology*, *93*:252–258 (1990).

[85] Vida, Stephen. "A Computer Program for Non-Parametric Receiver Operating Characteristic Analysis," *Computer Methods and Programs in Biomedicine*, *40*:95–101 (1993).

[86] Vida, Stephen. *AccuROC for Windows 95, Version 1.2*. Department of Psychiatry, McGill University Health Center, Montreal General Hospital, 1650 Cedar Avenue, Montreal, Quebec, H3G 1A4, Canada, January 1999.