UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS DA UNIVERSIDADE DE LISBOA
DEPARTAMENTO DE ESTATÍSTICA E INVESTIGAÇÃO OPERACIONAL

AS CURVAS ROC COMO INSTRUMENTO NA ANÁLISE ESTATÍSTICA DE TESTES DE DIAGNÓSTICO

CARINA SOARES DA SILVA

MESTRADO EM PROBABILIDADES E ESTATÍSTICA

2003

UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS DA UNIVERSIDADE DE LISBOA
DEPARTAMENTO DE ESTATÍSTICA E INVESTIGAÇÃO OPERACIONAL

AS CURVAS ROC COMO INSTRUMENTO NA ANÁLISE ESTATÍSTICA DE TESTES DE DIAGNÓSTICO

Dissertação orientada por:

Professora Doutora Maria Antónia Amaral Turkman

CARINA SOARES DA SILVA
MESTRADO EM PROBABILIDADES E ESTATÍSTICA
2003

Agradecimentos

No decorrer deste trabalho, tive a oportunidade de contar com o apoio de professores e amigos, a quem quero expressar o meu reconhecimento:

- À Profª Doutora Antónia Amaral Turkman que, em todos os momentos, me orientou e motivou nesta dissertação, e a quem devo o valioso contributo para a minha formação profissional
- Ao Dr Vinhas e à Técnica Leonor Campos, a total disponibilidade e prontidão na cedência de dados, mesmo quando tudo parecia impossivel
- À Profª Doutora Ana Cristina da Silva Braga, pela disponibilização do seu trabalho de doutoramento
- Ao Esteves e à Alexandra pela ajuda com o Latex
- Ao Rui Fortes por ter contribuido com a preciosa ajuda na conversão de ficheiros e pelo alento nos momentos mais difíceis
- Aos meus pais por me apoiarem

Resumo

Os testes de diagnóstico são bastante utilizados em muitas áreas da sociedade tecnológica moderna e visam utilizar métodos que permitam discriminar indivíduosentre duas populações. Têm particular interesse na medicina, uma vez que diagnósticos precoces e precisos podem fazer decrescer rácios de mortalidade e morbilidade (relação entre o número de casos de enfermidade e o número de habitantes). A necessidade de analisar a eficiciência destes testes é muito importante; um falso diagnóstico implica inapropriadas ou desnecessárias acções terapêuticas, que resultam em elevados custos financeiros e emocionais.

Uma vez que os testes de diagnóstico visam utilizar métodos que permitam discriminar entre populações de doentes e não doentes, a avaliação da exactidão de um teste de diagnóstico é imperativa. O custo associado a falsos negativos e a falsos positivos reflecte-se em primeiro lugar no indivíduo que realizou o referido teste.

A metodologia associada às curvas *Receiver Operating Characteristic* (ROC) permite avaliar a performance de um teste de diagnóstico, estabelecer o ponto de corte optimizando a relação entre a sensibilidade e a especificidade, comparar performances de vários testes de diagnótico, etc.

O objectivo deste trabalho é caracterizar as curvas ROC em termos de propriedades, métodos de estimação da curva, métodos para calcular a área abaixo da curva (AAC), cálculo do ponto de corte óptimo, métodos de comparação de testes de diagnóstico.

Palavras-Chave: Testes de diagnóstico, curvas ROC, sensibilidade, especificidade, métodos paramétricos e não-paramétricos, ponto de corte, área abaixo da curva (AAC), área parcial abaixo da curva (APAC).

Abstract

The diagnosis tests are widely used in several areas of the modern technological society and aim at to use methods that allow discrimination among individuals between two populations.

These tests have particular interest in medicine since precocious and precise diagnosis can decrease mortality and morbidity (relation between the number of disease cases and the number of inhabitants) ratios. The necessity to analyze the efficiency of these tests is very important; a false diagnosis implies inappropriate or unnecessary therapeutic actions that result in high financial and emotional costs.

Since diagnosis tests aim at to use methods that allow discrimination between sick and non sick populations, the evaluation of the exactness of a diagnosis test are imperative. The cost of false negatives and false positives is associated in the first place to the individual that carried through the test.

The methodology associated to the Receiver Operating Characteristic (ROC) allows the evaluation of the diagnosis test performance, to establish the cut point optimizing the relation between sensitivity and the specificity, to compare performances of several diagnosis tests, etc.

The purpose of this work is to characterize ROC curves in terms of properties, esteem methods of the curve, methods to calculate the area below of the curve (AUC), calculation of the optimum cut point, methods of comparison of diagnosis tests.

Key-Words: diagnosis tests, ROC curves, sensitivity, specificity, parametric and non-parametric methods, cut point, area below of the curve (AUC), partial area below of the curve (APAC).

Nota Introdutória

A análise ROC tem vindo a revelar-se uma ferramenta muito útil na avaliação do desempenho de testes de diagnóstico. Este trabalho vai abordar os testes de diagnóstico num contexto clínico. O termo testes de diagnóstico, geralmente aplica-se aos exames complementares de diagnóstico, no entanto devem ser entendidos num sentido mais amplo, abrangendo não só os exames complementares de diagnóstico como também os dados provenientes da história clínica e exame físico do indivíduo.

Uma vez que os testes de diagnóstico em geral visam utilizar métodos que permitam discriminar entre populações de doentes e não doentes, a avaliação da exactidão do teste em questão depende da comparação de resultados a partir dele obtidos com o verdadeiro estado de cada indivíduo. Para caracterizar correctamente este estado tem de se considerar um outro teste (ou conjunto de testes) cujo resultado seja determinante do estado do indivíduo com grande exactidão. A este teste dá-se o nome de *gold standard*.

A exactidão dos testes de diagnóstico é essencialmente avaliada por duas características: a sensibilidade, que se traduz na proporção de indivíduos doentes correctamente diagnosticados e pela especificidade, que designa a proporção de indivíduos não doentes correctamente diagnosticados. A curva ROC traduz-se numa representação gráfica de todos os valores da sensibilidade e da especificidade para um determinado teste de diagnóstico.

Num primeiro capítulo faremos uma introdução aos testes de diagnóstico e à descrição dos dados utilizados neste trabaho.

No capítulo dois, vamos iniciar a abordagem à metodologia ROC. Apresentamos a definição e propriedades da curva ROC. Depois apresentamos métodos de estimação paramétricos e não-paramétricos da curva ROC, ilustrados com exemplos de aplicação.

No capítulo três, vamos abordar várias técnicas paramétricas e não-paramétricas e a respectiva comparação, para calcular um dos índices mais utilizados para avaliar a performance de um teste de diagnóstico, a área abaixo da curva

(AAC) ROC..

No capítulo quatro, apresentam-se vários métodos para comparar testes de diagnóstico, quando temos amostras emparelhadas, independentes e parcialmente emparelhadas. Também abordamos a estimação da área parcial abaixo da curva (APAC) e a comparação das APACs em testes de diagnóstico.

No capítulo cinco pretende-se abordar técnicas para o cálculo do ponto de corte óptimo.

No capítulo seis faremos as considerações finais e conclusões. Também será abordada uma nova metodologia para avaliar a performance de testes de dignóstico que discriminam em três populações a chamada Superficie ROC.

No desenvolvimento deste trabalho recorreu-se a várias aplicações informáticas: SPSS, Excell, subrotinas do FORTRAN, ROCKIT e PROPROC.

Conteúdo

Lista de Figuras

Lista de Tabelas

Capítulo 1

Testes de Diagnóstico

1.1 Introdução

Os testes de diagnóstico são instrumentos importantes para tomar decisões que os médicos são obrigados a realizar como parte intrínseca da sua actividade. Geralmente obedecem a vários tipos de decisões, como por exemplo: confirmar a presença de uma doença; avaliar a gravidade do quadro clínico; estimar o prognóstico de uma doença; avaliar a resposta de uma conduta terapêutica.

O teste de diagnóstico ideal daria sempre respostas correctas - positivo para a presença da doença e negativo para a ausência -, seria rápido, seguro, íncruento e barato. Mas, na prática, não existem testes ideais.

Os testes de diagnóstico são classificados em dois tipos:

- *Testes Qualitativos*: A variável resposta (variável que representa os valores do teste de diagnóstico) é dicotómica, o teste tem um resultado positivo ou negativo (presença ou ausência da doença). Por exemplo, a biópsia miocárdica com resultado positivo ou negativo para a re-

jeição do transplante cardíaco. Ou então a variável resposta é ordinal (*rating data*). Por exemplo em Radiologia é comum usar uma escala ordinal de cinco pontos: 1=doença definitavemente ausente, 2=doença provavelmente ausente, 3=doença possivelmente presente, 4=doença provavelmente presente, 5=doença definitivamente presente [14].

- *Testes Quantitativos*: A variável resposta é contínua e a classificação dos pacientes é dada a partir de um ponto de corte seleccionado de acordo com um critério adequado.

Os testes, para serem úteis, devem ser precisos e exactos. Um teste preciso é aquele que mantém a reprodutibilidade em várias medições sucessivas na mesma amostra e um teste exacto é aquele que revela em geral um valor do teste igual ao verdadeiro valor [6]. A figura 1.1 demonstra a relação entre um teste exacto e preciso.

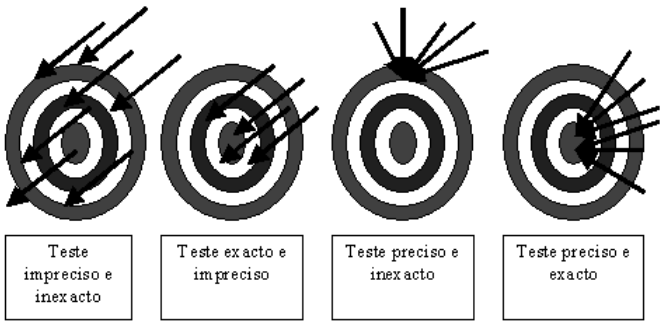

Figura 1.1: Relação entre um teste exacto e preciso

Para que um teste de diagnóstico possa ser utilizado em indivíduos futuros é preciso que ele seja considerado adequado. Assim para estudar a qualidade de um teste de diagnóstico em termos da sua exactidão e precisão é necessário fazer uma experiência com indivíduos para os quais se conhece o seu verdadeiro estado. Este conhecimento é obtido através de um teste com uma grande exactidão e precisão, que se denomina por *gold standard*, mas que em geral é dispendioso, invasivo e não é recomendável a sua aplicação indiscriminada. Quando não existe um *gold standard*, pode-se utilizar o mel-

hor teste de referência disponível.

A relação entre o *gold stdandard* e um teste de diagnóstico pode ser sumariada na tabela 1.1:

	Doença Presente (D)	**Doença Ausente ($\bar{D}$)**
Teste Positivo (+)	Verdadeiro Positivo (VP)	Falso Positivo (FP)
Teste Negativo (-)	Falso Negativo (FN)	Verdadeiro Negativo (VN)

Tabela 1.1: Relação do *gold standard* com o teste de diagnóstico

Se se tiver um total de n indivíduos, dos quais sabemos o resultado do teste (positivo ou negativo) e o seu verdadeiro estado (doente ou não doente), podemos resumir esses dados na tabela 1.2.

	Doença Presente (D)	**Doença Ausente ($\bar{D}$)**	Total
Teste Positivo (+)	a	b	$a+b$
Teste Negativo (-)	c	d	$c+d$
Total	$a+c$	$b+d$	n

Tabela 1.2: Tabela de Contingência

A exactidão dos testes de diagnóstico é essencialmente avaliada por duas características: a *sensibilidade* (S) e a *especificidade* (E). Probabilisticamente define-se a sensibilidade como a probabilidade condicional de um teste positivo dado que o indivíduo é doente, ou seja o teste classifica correctamente o indivíduo doente; e a especificidade é a probabilidade condicional de um teste negativo dado que o indivíduo é não doente, ou seja o teste classifica correctamente o indivíduo não doente.

Para simplificar a notação passamos a representar simbolicamente os acontecimentos na tabela 1.3:

Acontecimento	**Representação simbólica**
O indivíduo é doente	D
O indivíduo é não doente	$\bar{D}$
O teste é positivo	+
O teste é negativo	-

Tabela 1.3: Notação para os acontecimentos

Consequentemente representamos as probabilidades relevantes na tabela 1.4.

Sigla	**Probabilidade**	**Representação simbólica**
PVP ou S	Probabilidade de um indivíduo ter teste positivo dado que é doente	$P(+\|D)$
PVN ou E	Probabilidade de um indivíduo ter teste negativo dado que é não doente	$P(-\|\bar{D})$
PFN	Probabilidade de um indivíduo ter teste negativo dado que é doente	$P(-\|D)$
PFP	Probabilidade de um indivíduo ter teste positivo dado que é não doente	$P(+\|\bar{D})$
VPP	Probabilidade de um indivíduo ser doente dado que o teste é positivo	$P(D\|+)$
VPN	Probabilidade de um indivíduo ser não doente dado que o teste é negativo	$P(\bar{D}\|-)$
P	Prevalência	$P(D)$
E_f	Probabilidade de um indivíduo ter o resultado do teste verdadeiro	$P(EF)$
	Probabilidade de um indivíduo ter teste positivo	$P(+)$
	Probabilidade de um indivíduo ter teste negativo	$P(-)$
	Probabilidade de um indivíduo ser doente e ter teste positivo	$P(D,+)$
	Probabilidade de um indivíduo ser não doente e ter teste positivo	$P(\bar{D},+)$
	Probabilidade de um indivíduo ser não doente e ter teste negativo	$P(\bar{D},-)$
	Probabilidade de um indivíduo ser doente e ter teste negativo	$P(D,-)$
	Probabilidade de um indivíduo ser verdadeiro positivo	$P(VP)$
	Probabilidade de um indivíduo ser verdadeiro negativo	$P(VN)$
	Probabilidade de um indivíduo ser falso positivo	$P(FP)$
	Probabilidade de um indivíduo ser falso negativo	$P(FN)$

Tabela 1.4: Notação para as probabilidades

Então a partir da tabela 1.2, podemos estimar estas probabilidades através das frequências de ocorrência:

$$\widehat{S} = \frac{a}{a+c} \qquad \text{e} \qquad \widehat{E} = \frac{d}{b+d}$$

A sensibilidade designa-se também por proporção de verdadeiros positivos (PVP), a especificidade por proporção de verdadeiros negativos (PVN).

Para além destas características podemos também estimar:

- PFN pela proporção de respostas negativas no grupo de doentes:

$$\widehat{PFN} = \frac{c}{a+c}$$

  ou seja $1 - \widehat{S}$

- PFP pela proporção de respostas positivas no grupo de não doentes:

$$\widehat{PFP} = \frac{b}{b+d}$$

  ou seja $1 - E$

- Valor preditivo positivo,VPP, pela proporção de indivíduos com o teste positivo que são doentes, também denominada por probabilidade pós-teste:

$$\widehat{VPP} = \frac{a}{a+b}$$

- Valor preditivo negativo, VPN, pela proporção de indivíduos com teste negativo que são não doentes:

$$\widehat{VPN} = \frac{d}{c+d}$$

- Prevalência (P) pela proporção de indivíduos doentes, independentemente do resultado do teste de diagnóstico, também denominada de probabilidade pré-teste:

$$\widehat{P} = \frac{a+c}{n}$$

  Por vezes esta probabilidade é conhecida antes de se realizar o teste.

- Exactidão ou eficiência global do teste (E_f) pela proporção de indivíduos que tiveram o resultado do teste de diagnóstico verdadeiro:

$$\widehat{E_f} = \frac{a+d}{n}$$

Estas quantidades estimam empiricamente as probabilidades correspondentes.

A partir do Teorema de Bayes, podemos estabelecer as seguintes relações entre algumas variáveis:

- $$VPP = \frac{S \times P}{S \times P + (1 - E) \times (1 - P)}$$

  Demonstração:

$$VPP = P(D|+) = \frac{P(D,+)}{P(+)} = \frac{P(D) \times P(+|D)}{P(D,+) + P(\bar{D},+)} =$$
$$= \frac{P(D) \times P(+|D)}{P(D) \times P(+|D) + P(\bar{D}) \times P(+|\bar{D})} =$$
$$= \frac{P \times S}{P \times S + (1 - P) \times (1 - E)}$$

◁

- $$VPN = \frac{E \times (1 - P)}{E \times (1 - P) + (1 - S) \times P}$$

  Demonstração:

$$VPN = P(\bar{D}|-) = \frac{P(\bar{D},-)}{P(-)} =$$
$$= \frac{P(\bar{D}) \times P(-|\bar{D})}{P(D,-) + P(\bar{D},-)} = \frac{P(\bar{D}) \times P(-|\bar{D})}{P(D) \times P(-|D) + P(\bar{D}) \times P(-|\bar{D})} =$$
$$= \frac{(1 - P) \times E}{P \times (1 - S) + (1 - P) \times E}$$

◁

- $$E = 1 - PFP$$

  Demonstração:

$$E = P(-|\bar{D}) = \frac{P(- \cap \bar{D})}{P(\bar{D})} = \frac{P(\bar{D}) \times P(-|\bar{D})}{P(\bar{D})} =$$
$$= \frac{P(\bar{D}) \times (1 - P(+|\bar{D}))}{P(\bar{D})} = 1 - \frac{P(\bar{D}) \times P(+|\bar{D})}{P(\bar{D})} =$$
$$= 1 - \frac{P(\bar{D} \cap +)}{P(\bar{D})} = 1 - P(+|\bar{D}) = 1 - PFP$$

◁

A avaliação de um teste de diagnóstico depende essencialmente da sensibilidade e da especificidade. Estas caracteristicas não dependem da prevalência da doença. Mas, quanto maior a sensibilidade maior será o valor preditivo negativo, ou seja, maior será a probabilidade de, perante um resultado negativo, não haver doença. Quanto maior a especificidade, maior será o valor preditivo positivo, isto é, maior será a probabilidade de, perante um resultado positivo, haver doença. Quanto maior for a prevalência, maior será o valor preditivo positivo e menor o valor preditivo negativo, ou seja, quanto mais frequente for a doença mais provável é encontrar verdadeiros positivos (aumentando o valor preditivo positivo), mas também é mais provável encontrar falsos negativos (diminuindo o valor preditivo negativo).

Testes sensíveis são aqueles que geralmente são positivos quando a doença está presente. Se todos os indivíduos portadores de uma determinada doença apresentam resultados positivos, esse teste tem uma sensibilidade de 100 %. Testes sensíveis são úteis quando existe uma penalização importante para a omissão do diagnóstico; em programas de rastreio; no início da avaliação de um doente, quando estão a ser consideradas muitas possibilidades de diagnóstico. Podemos ter testes específicos, aqueles que geralmente são negativos

quando a doença está ausente. Se todos os indivíduos não doentes apresentam um teste negativo, este teste tem uma especificidade de 100 %. Este tipo de teste é útil, por exemplo, quando se pretende confirmar um diagnóstico que é sugerido por testes menos específicos; quando a existência de um falso positivo tem importantes implicações físicas, emocionais e financeiras para o paciente.

Quando estamos na presença de um teste quantitativo, há necessidade de decidir qual o valor do teste a partir do qual se deve considerar que o indivíduo é ou não doente. Este ponto é chamado de ponto de corte. O "eixo de decisão" define-se como sendo o conjunto de todos os valores possíveis que o teste de diagnóstico pode tomar. Os investigadores devem avaliar cuidadosamente a sensibilidade e a especifidade do teste para estabelecer o ponto de corte mais adequado. Quando o ponto de corte é movido no eixo de decisão, um aumento da sensibilidade resulta no sacrifício da especificidade. Podemos mover o ponto de corte em ambas as direcções do eixo de decisão, dependendo se o investigador preferir um teste mais sensível ou mais específico. A localização ideal de um ponto de corte depende de um equilibrio adequado entre a sensibilidade e a epecificidade e não é simples a selecção da combinação óptima entre a sensibilidade e a especificidade.

Podemos fazer uma analogia com os testes de hipóteses em Estatística. Quando exigimos uma probabilidade menor para o "erro de tipo I (α)", que consiste em rejeitar uma hipótese nula verdadeira, resulta num acréscimo do "erro de tipo II (β)", que consiste em não rejeitar a hipótese nula quando a alternativa é verdadeira, "*Não se pode ter o melhor de dois mundos, há que fazer opções.*"[35].

A escolha da sensibilidade e da especificidade depende da natureza da doença, da população clínica e do custo relativo de falsos positivos e falsos negativos. Para uma doença rara, ter uma especificidade elevada é importante, no sentido de evitar um grande número de falsos positivos. Para uma doença mais comum, uma elevada especificidade não é tão importante. Por exemplo a sensibilidade que usualmente se aplica para detectar um cancro do pulmão é de 70% com uma especificidade de 99%, quanto às mamografias a sensibilidade está entre 70% e 80% e a especificidade é cerca de 95% [1].

1.2 Testes de Diagnóstico para detecção de alergias

Quinze por cento da população mundial sofre de reacções alérgicas. Somente na Europa e E.U.A., existem mais de 50 milhões de doentes alérgicos (*The UCB Institute of Allergy*). Algumas pessoas são muito sensíveis ao contacto com matéria vegetal (plantas, frutos, vegetais) ou com determinados alimentos (peixe, marisco) ou, mesmo, com alergenos presentes na atmosfera (poeira, pólen, detritos de ácaros, pêlos de animais). O contacto desencadeia uma reacção no organismo e pode provocar queixas respiratórias, oculares, digestivas e cutâneas. Estas reacções excessivas do organismo são designadas por "alergias". A pergunta óbvia é: porque razão algumas pessoas as têm e outras não? A resposta reside em conhecer o mecanismo da alergia.

A matéria vegetal e os alimentos, a poeira e o pólen são estranhos ao corpo humano e denominados de antigénios. Quando os antigénios entram em contacto com o corpo humano, este defende-se através da libertação de anticorpos, produzidos pelos glóbulos brancos, que ajudam a combater a "invasão" dos antigénios. O problema é que em certas pessoas (aproximadamente 15% da população mundial) o combate entre os antigénios e os anticorpos provoca uma reacção excessiva, ou "alérgica". Neste caso, os antigénios denominam-se de alergenos.

Os sintomas de alergia são um resultado de acontecimentos que se realizam no sistema imunológico, que é o mecanismo de defesa do organismo contra substâncias estranhas. O organismo identifica certas substâncias, chamadas de alergenos, como estranhos. Estas substâncias que são inofensivas para a maior parte das pessoas, desencadeiam reacções alégicas dentro do sistemas imunológico da pessoa.

Quando alguém com predisposição para alergias e encontra um alergeno que ao qual é sensível, desencadeiam-se vários acontecimentos. O principal culpado que desencadeia estes acontecimentos é um anticorpo chamado de Imunoglobulina E (IgE). A IgE "defende" o corpo procurando os alergenos ofensivos dos tecidos e do fluxo sanguíneo. A primeira vez que entra um alergeno no corpo de uma pessoa alérgica, este reage produzindo anticorpos

de IgE. Estes anticorpos instalam-se nas células chamadas mastócitos e esperam a próxima vez que os alergenos entrem no organismo. Quando entram, os anticorpos de IgE capturam os alergenos e elimina-os da circulação. Os mastócitos reagem libertando agentes químicos especiais chamados de "mediadores". Estes mediadores produzem os sintomas clássicos das reacções alérgicas.

Sem um diagnóstico e tratamento adequados as condições do paciente pioram com o tempo. A princípio é muito difícil diagnosticar uma alergia. Antes do mais, o médico tem que verificar, entre os sintomas, quais os que correspondem aos da alergia e, depois, identificar o alergeno que causa o problema. Uma vez convencido de que se encontra na pista certa, o médico executa diversos testes para verificar se o seu primeiro diagnóstico está correcto. De entre estes testes figuram, os da pele, destinados a avaliar o grau de sensibilidade do doente a uma série de alergenos que o médico suspeita que sejam os culpados. O teste consiste em colocar o organismo em contacto directo com uma quantidade mínima do alergeno e em observar a reacção. É realizada uma lesão superficial na pele da face volar do antebraço, sem provocar sangramento. É colocada uma gota de alergeno e, após sensivelmente vinte minutos, é realizada a leitura do resultado medindo os diâmetros tranversais da pápula. São considerados testes cutâneos positivos, aqueles que determinam a formação de uma pápula superior ou igual a 3 mm [39].

Outro teste é a IgE Total que determina a quantidade total de anticorpos IgE no soro ou plasma humano. A IgE é produzida como resultado da exposição repetida ao alergeno. O nível de IgE no soro varia com a idade, a tabela 1.5 indica os valores de referência por faixa etária.

A IgE Total encontra-se mais elevada na presença de infecções alérgicas, permancendo com valores superiores aos dos limites de referência.

Outro teste de diagnóstico é o RAST. Este teste é realizado quando não se podem fazer testes cutâneos. É realizada uma colheita de sangue e é exposta a diferentes alergenos. Anticorpos de IgE "atacam" os alergenos se se for alérgico. Um químico radioactivo é utilizado para fazer a análise de resultados.

Idade	Valores
de 1 a 12 meses	0.00-11.0
até 2 anos	0.00-11.00
de 2 a 3 anos	0.00-15.00
de 3 a 5 anos	0.00-39.00
de 5 a 8 anos	0.00-79.00
de 8 a 12 anos	0.00-208.00
de 12 a 16 anos	0.00-106.00
desde os 16 anos	0.00-68.00

Tabela 1.5: Valores de referência por faixa etária divulgados pelo laboratório: Laboratório Domingo

Descrição dos dados

A amostra é constituída por 111 indivíduos com idade superior a 15 anos e de ambos os sexos. As variáveis observadas foram a idade, o sexo, valores do teste de diagnóstico de IgE Total e o teste *gold standard* que resultou da análise de dois testes de diagnóstico, os cutâneos e o RAST, o alergologista ao analisar estes dois testes classifica o indivíduo como doente ou não doente. A tabela 1.6 caracteriza as variáveis em estudo

Caracterização do Teste de Diagnóstico IgE Total

A amostra foi dividida em duas sub-amostras, doentes e não doentes de acordo com o *gold standard*. Na tabela 1.7 apresentam-se algumas estatísticas para a variável IgE Total para as duas amostras e as figuras 1.2 e 1.3 representam as respectivas distribuições empíricas.

Variável	**Domínio**	**Escala**	**Código**
Idade	$>$ 15 anos	Métrica	Idade
Sexo	1=masculino 0=feminino	Nominal	sexo
IgE Total	$>$0	Métrica	IgETotal
gold standard	1=positivo 0=negativo	Nominal	gs

Tabela 1.6: Varíáveis em estudo

	Doentes	**Não Doentes**
Dimensão	51	60
Média	424.3231	109.8805
Desvio Padrão	793.3695	149.3031
Mínimo	13	4.10
Máximo	5000	626
Coeficiente de Assimetria	4.514	2.022

Tabela 1.7: Caracterização da amostra

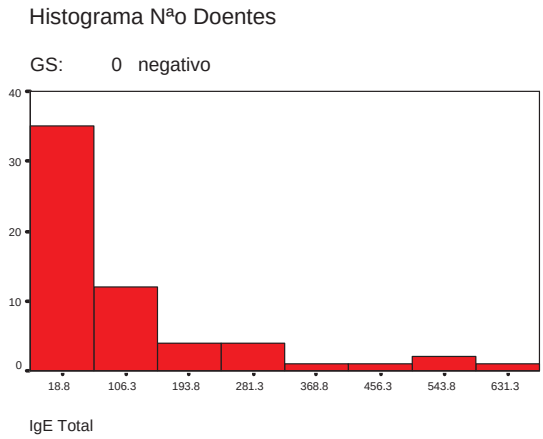

Figura 1.2: Histograma Não Doentes

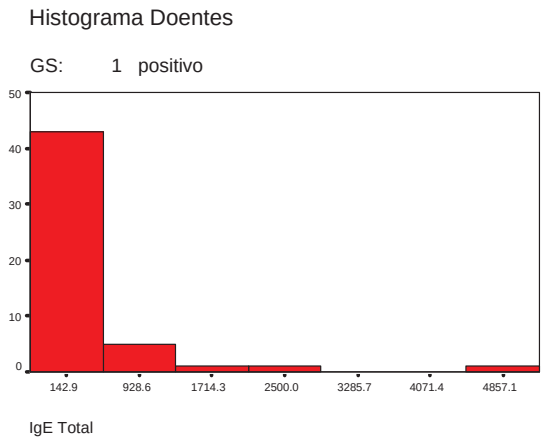

Figura 1.3: Histograma Doentes

Quer pelos coeficientes de assimetria (tabela 1.7) quer pelos histogramas (figuras 1.2 e 1.3), podemos concluir que se tratam de duas amostras enviesadas positivamente. Por forma a minimizar este enviesamento procedeu-se a uma transformação logarítmica dos dados e fomos caracterizá-los (tabela 1.8 e histogramas 1.4 e 1.5).

	Doentes	**Não Doentes**
Dimensão	51	60
Média	5.33	3.85
Desvio Padrão	1.163	1.41
Mínimo	2.56	1.41
Máximo	8.52	6.44
Coeficiente de Assimetria	0.269	0.046

Tabela 1.8: Caracterização do logaritmo das amostras

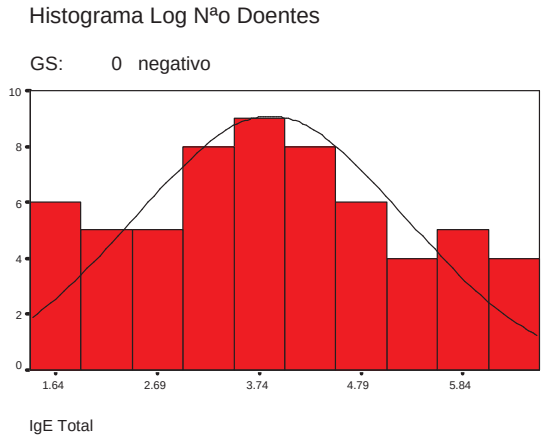

Figura 1.4: Histograma logaritmo Não Doentes

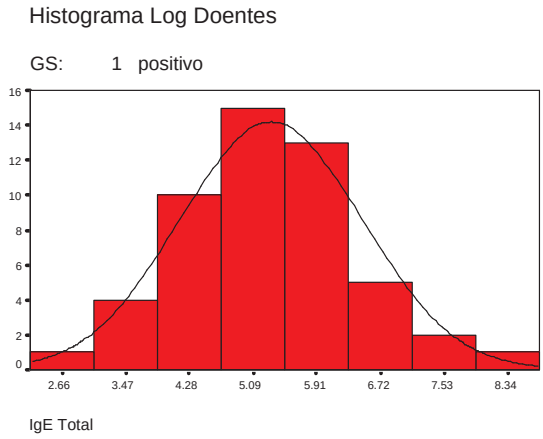

Figura 1.5: Histograma logaritmo Doentes

Como podemos observar a transformação logarítmica levou-nos a minimizar o enviesamento nas duas amostras. Fomos testar o ajustamento à normalidade dos dados a partir do teste de *Kolmogorvov-Smirnov*, e para um valor-p de 0.2, os dados seguem uma distribuição Gaussiana.

Capítulo 2

Curva ROC

2.1 Introdução

A análise ROC baseia-se na Teoria de Detecção de Sinais, desenvolvida durante a 2^a Guerra Mundial, com o objectivo de analisar imagens de radares. O operador de radar tinha de decidir se um ponto no écran representava um alvo inimigo, um navio amigável ou simplesmente ruído [32]. Os operadores eram denominados por *Radar Receiver Operators*. A Teoria de Detecção de Sinais tinha por objectivo avaliar a performance dos operadores (*Receiver Operating Characteristic*).

O conceito de *Receiver Operating Characteristic Curves* foi introduzido na medicina por Lusted nos finais da década de 60 [26]. Na década de 70 assistiu-se a um desenvolvimento significativo na metodologia estatística em relação a diagnósticos médicos. Este desenvolvimento deveu-se em grande parte à área da epidemiologia, visto que influenciou a selecção de alguns tópicos que reservaram maior atenção à identificação e correcção de erros, e menor atenção a tópicos técnicos tais como a análise discriminante. Esta tendência tem a ver com o aumento do número de investigadores médicos que também possuem conhecimentos computacionais e quantitativos. Também

se reflectiu no desenvolvimento de jornais com uma interface bioestatística e clínica, como por exemplo *Statistics in Medicine* [3].

Swets and Pickett [40] enumeram quase duzentas áreas de aplicação das curvas ROC. Vão desde a Teoria de Detecção de Sinais, Psicologia, Detecção Poligráfica de Mentiras, Epidemiologia, Nutrição, Imagiologia, Psiquiatria, Inspecção de Sistemas Fabris, etc. Podemos aplicar a análise ROC a qualquer área desde que se pretenda discriminar duas populações.
Em muitas aplicações estatísticas estamos interessados em classificar uma determinada unidade num de dois grupos. Por exemplo, podemos querer classificar indivíduos em fraco ou em alto risco de enfarte, baseando-nos na medição da pressão sanguínea dos indivíduos. A discriminação em dois grupos é feita a partir de um ponto de corte a definir. Esta discriminação pode não ser perfeita, de acordo com o exemplo anterior, pode haver indivíduos que tenham uma pressão sanguínea relativamente elevada e no entanto não estarem em risco de enfarte.

A partir do ponto de corte calculam-se a sensibilidade e a especificidade do teste de diagnóstico. Sempre que se altere o critério de diagnóstico (ou seja o ponto de corte) os valores da sensibilidade e da especificidade também se alteram. Se num plano unitário representarmos todos os valores da sensibilidade *vs* (1-especificidade), para todos os pontos de corte possíveis no eixo de decisão, obtemos a chamada curva ROC 2.1.

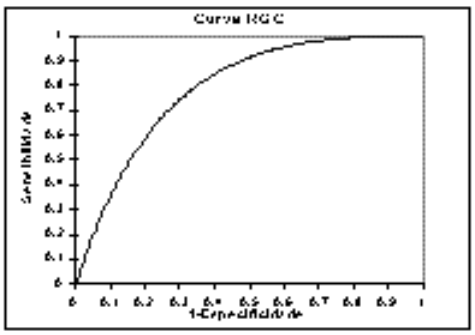

Figura 2.1: Representação da Curva ROC no plano unitário

Para além da avaliação de um teste de diagnóstico e a escolha do ponto de corte óptimo, também podemos usar as curvas ROC para comparar vários testes de diagnóstico.

2.2 Definição e propriedades da curva ROC

Suponhamos que temos duas populações de indivíduos, uma correspondente a indivíduos não doentes e outra de indivíduos doentes. Seja f_X a função de densidade de probabilidade (f.d.p.) associada aos valores do teste de diagnóstico dos indivíduos não doentes e g_Y a f.d.p. associada aos valores do teste de diagnóstico dos indivíduos doentes. F_X e G_Y são as correspondentes funções de distribuição. A cada ponto de corte c está associada uma regra de classificação binária que define se o teste é positivo ou negativo. Sem perda de generalidade, assume-se que se a variável resposta é superior a c, o teste é positivo e se a variável resposta é inferior a c, o teste é negativo. O ponto de corte c dá origem a duas áreas abaixo da curva de densidade de probabilidade das funções f_X e g_Y, como estão designadas na figura 2.2.

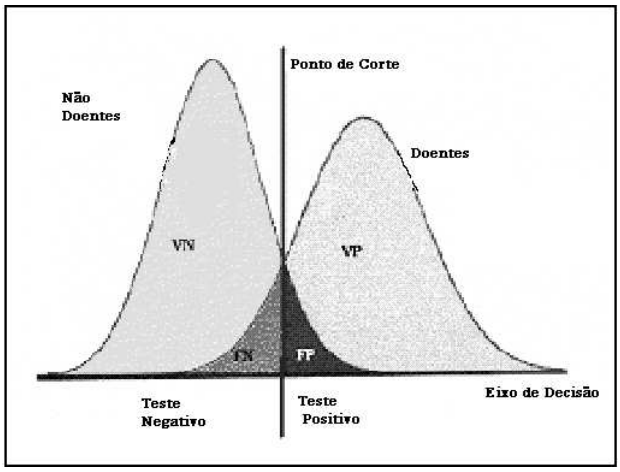

Figura 2.2: Funções de densidade hipotéticas para cada variável resposta dos indivíduos doentes e não doentes

Para a direita do ponto de corte (teste positivo) identificamos uma área

correpondente aos falsos positivos (FP) e outra aos verdadeiros positivos (VP), para a esquerda do ponto de corte (teste negativo) identificamos uma área correspondente aos verdadeiros negativos (VN) e outra aos falsos negativos (FN). Quanto menor for a sobreposição das distribuições, menor é a área correspondente aos falsos positivos e falsos negativos. A razão $\frac{S}{1-E}$ decresce para um à medida que c decresce. Um valor baixo de c conduz a um baixo grau de confiança na hipótese de que um valor do teste de diagnóstico maior ou igual a c esteja associado à população dos doentes.

O desejável seria ter um teste de diagnóstico com 100 % de sensibilidade e 100 % de especificidade. Isso é de esperar de um *gold standard*! Isso não acontece em testes de diagnóstico que se querem simples, de baixo custo e rápidos. Em geral o que acontece é um aumento da sensibilidade implicar um decréscimo da especificidade e vice-versa. Este comportamento é fácil de observar se se associar o aumento da especificidade a movimentos do ponto de corte para a direita (figura 2.3) e o aumento da sensibilidade a movimentos do ponto de corte para a esquerda (figura 2.4).

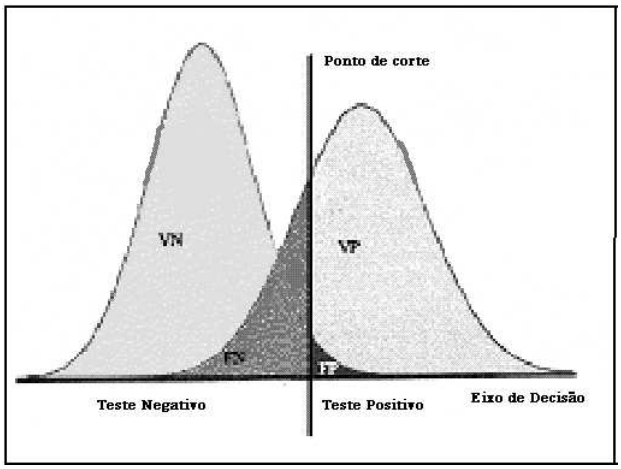

Figura 2.3: Movimento do ponto de corte para a direita

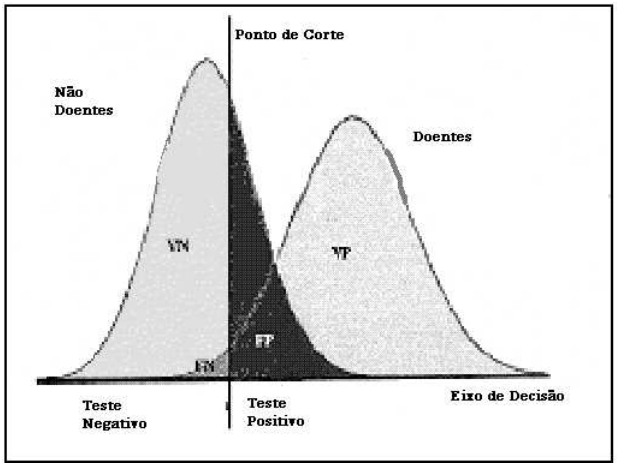

Figura 2.4: Movimento do ponto de corte para a esquerda

As curva ROC permitem estabelecer um ponto de corte de forma a optimizar a relação sensibilidade *vs* especificidade.

Definição de Curva ROC

Vamos definir a especificidade (E) e a sensibilidade (S) por:

$$E = F(c) = P(X \leq c)$$
$$S = 1 - G(c) = 1 - P(Y \leq c)$$

Schäfer [20] define a curva ROC do seguinte modo:

$$\varphi(E) = 1 - G(F^{-1}(E)), \qquad 0 \leq E \leq 1 \tag{2.1}$$

onde F^{-1} é a função inversa de F, e o quantil de probabilidade E da função F é definido do seguinte modo:

$$F^{-1}(E) = inf\{x \in S(F) : F(x) \geq E\} \qquad 0 \leq E \leq 1 \tag{2.2}$$
$$\text{onde} \qquad S(F) = \{x \in \mathbf{R} : 0 < F(x) < 1\} \qquad \text{é o suporte de } F.$$

A curva ROC resulta da representação gráfica de todos os pares (1-E, S) no plano unitário para todos os possíveis pontos de corte [14].

Alguns autores usam as escalas invertidas e definem a *ordinal dominance curve* (ODC), que resulta em representar os pontos (E, 1-S) [19], que possuem as mesmas características que as curvas ROC.

Propriedades da curva ROC

1. A curva ROC é invariante em relação a uma transformação monótona da escala de X e Y

Demonstração:

A curva ROC é definida pela função:

$$\varphi : [0,1] \longrightarrow [0,1]$$
$$\varphi(E) = 1 - G(F^{-1}(E))$$

e pretende-se provar que $\forall E \in [0,1], \varphi(E) = \varphi_1(E)$

Se X for uma variável aleatória com função de distribuição $F(x)$ e Y uma variável aleatória com função de distribuição $G(y)$, e $h(.)$ uma transformação monótona, então:

$$\begin{aligned} X_1 &= h(X) \\ Y_1 &= h(Y) \end{aligned}$$

$$F(x) = P(X \leq x) = P(h(X) \leq h(x)) = P(X_1 \leq h(x)) = F_1(h(x))$$

$$G(y) = P(Y \leq y) = P(h(Y) \leq h(y)) = P(Y_1 \leq h(y) = G_1(h(y))$$

isto é, $F(x) = F_1(h(x))$ e $G(y) = G_1(h(y))$.

Com efeito por (2.1) tem-se

$$\varphi_1(E) = 1 - G_1(F_1^{-1}(E))$$

Por outro lado tem-se

$$h(x) = F_1^{-1}(F(x))$$

e para $x = F^{-1}(E)$ obtém-se

$$h(F^{-1}(E)) = F_1^{-1}(E)$$

Consequentemente vem:

$$1 - G_1(h(F^{-1}(E)) = 1 - G(F^{-1}(E)) = \varphi(E)$$

◁

2. X é estocasticamente inferior a Y, isto é, $F_X(c) \geq G_Y(c) \quad \forall c$. Como consequência faz com que a curva ROC esteja definida acima da diagonal positiva do plano unitário, ou seja $\int_0^1 \varphi(E)dE \geq 0.5$

Demonstração:

$$\int_0^1 \varphi(E)dE = \int_0^1 [1 - G(F^{-1}(E))]dE =$$
$$= 1 - \int_0^1 G(F^{-1}(E))dE$$

Sabendo que

$$\begin{aligned} E &= F(c) \\ F^{-1}(E) &= c \\ dE &= f(c)dc \end{aligned}$$

vem que

$$1 - \int_{-\infty}^{+\infty} G(c)f(c)dc$$

Como $F_X(c) \geq G_Y(c)$, então

$$1 - \int_{-\infty}^{+\infty} G(c)f(c)dc \geq 1 - \int_{-\infty}^{+\infty} F(c)f(c)dc = 1 - \int_{-\infty}^{+\infty} F(c)F^{'}(c)dc =$$

$$= 1 - [\frac{F^2(c)}{2}]_{-\infty}^{+\infty} =$$

$$= 1 - [\frac{F^2(+\infty)}{2} - \frac{F^2(-\infty)}{2}] = 1 - [\frac{1}{2} - 0] = 0.5$$

e pela regra de transitividade $\int_0^1 \varphi(E)dE \geq 0.5$

◁

3. Se as densidades f e g tem uma razão de verosimilhanças monótona, então as curvas são côncavas

4. A área abaixo da curva é dada por $P(X < Y)$ isto é

$$\int_0^1 \varphi(E)dE = P(X < Y)$$

Demonstração:

$$\int_0^1 \varphi(E)dE = \int_0^1 (1 - G(F^{-1}(E))dE =$$

$$= \int_{-\infty}^{+\infty} (1 - G(c))dF(c) = \int_{-\infty}^{+\infty} (1 - \int_{-\infty}^{c} g(y)dy)dF(c) =$$

$$= 1 - \int_{-\infty}^{+\infty} \int_{-\infty}^{c} g(y)dyf(c)dc = 1 - \int_{-\infty}^{+\infty} P(Y \leq y)f(c)dc$$

e

$$P(X < Y) = 1 - P(X \geq Y) = 1 - \int_{-\infty}^{+\infty} \int_{-\infty}^{c} f(c, z)dcdz =$$

$$= 1 - \int_{-\infty}^{+\infty} \int_{-\infty}^{c} f_X(c)g_Y(z)dcdz = 1 - \int_{-\infty}^{-\infty} f_X(c)(\int_{+\infty}^{c} g_Y(z)dz)dc =$$

$$= 1 - \int_{-\infty}^{+\infty} P(Y \leq c)f_X(c)dc$$

Então

$$\int_0^1 \varphi(E)dE = P(X < Y)$$

◁

2.3 Métodos de Estimação da Curva ROC

A variável resposta pode ser de três tipos: binária (positivo ou negativo), as probabilidades correspondentes à especificidade e sensibilidade são directamente estimadas; definida numa escala ordinal (*rating data*) e contínua. A natureza dos dados, obviamente, condiciona a escolha do modelo de estimação da curva.

Há, essencialmente, dois tipos de métodos não-paramétricos, o empírico e o *kernel*. O empírico, talvez o mais vulgar, consiste em representar todos os pares (1-E, S) para todos os possíveis pontos de corte. Este método não necessita de postular nenhuma condição à variável resposta, quer para as populações dos indivíduos doentes quer dos não doentes. Uma das vantagens deste método é o facto de ser robusto, pois é livre de pressupostos distribucionais, e existem vários programas que o implementam (por exemplo SPSS) [32, 13, 48, 19]. Uma desvantagem deste método é produzir curvas irregulares, resultando numa subestimação da área abaixo da curva.

Para o método *kernel* [37] também não é estabelecido nenhum pressuposto relativamente às distribuições subjacentes aos dados (de acordo com a própria definição de não-paramétrico). Este método constrói uma curva suave a partir da estimação das funções de densidade dos indivíduos doentes e não doentes. A desvantagem deste método é a selecção da função *kernel* e a correspondente amplitude (*bandwith*), que requerem cálculos complexos e morosos.

Até 1980 os métodos paramétricos para ajustar uma curva ROC eram baseados unicamente no modelo Binormal, que estava intimamente relacionado com o facto de que inicialmente a análise ROC estava confinada a interpretar testes cuja variável resposta era em escala ordinal [14, 15, 3]. No entanto a resposta à pergunta "será que o modelo binormal se aplica a testes de diagnóstico com variável resposta contínua?"Metz *et al* [32] e outros [16] deram resposta esta questão . Metz ajusta uma curva ROC a dados contínuos. Primeiramente, discretiza os dados em tantas categorias quanto possível e depois utiliza o procedimento ROCFIT para obter estimativas de máxima verosimilhança para os dois parâmetros da binormal. Este

método é considerado semi-paramétrico e baseia-se no pressuposto de que as distribuições dos indivíduos doentes e não doentes são Gaussianas, ou então a partir de transformações monótonas (uma vez que a curva é invariante a transformações monótonas) da variável resposta para cada uma das populações. No entanto, em algumas situações não será o método mais adequado, como por exemplo quando as amostras são muito pequenas ou quando se fica com poucas categorias. Estas situações conduzem a curvas com "ganchos" nos cantos superior direito e inferior esquerdo do plano unitário. O modelo binormal "próprio"[28] é um método alternativo ao binormal convencional nestas situações. Metz *et al* [32] criaram dois algoritmos para estimar a função máxima de verosimilhança da curva ROC binormal para dados de natureza contínua, LABROC4 e LABROC5. O modelo binormal "Próprio" é uma alternativa ao modelo binormal convencional, quando se tem amostras de pequenas dimensões, ou quando os dados são de natureza categórica numa escala ordinal, e as observações não estão bem distribuídas pelas categorias, que leva a curvas ROC degeneradas.

Testes de diagnóstico com variáveis resposta contínuas lidam com uma grande variedade de formas de distribuição. É comum que as distribuições associadas às populações de indivíduos doentes sejam muito assimétricas enquanto que as associadas às populações dos indivíduos não doentes tendem a ser mais simétricas [38]. Uma transformação de escala poderia ser apropriado mas também pode introduzir distorção na avaliação do teste de diagnóstico. A distribuição_S é um método paramétrico para estimar curvas ROC para testes de diagnóstico cuja variável resposta é contínua. A distribuição_S é definida por uma equação diferencial onde a variável dependente é cumulativa. Este método providencia inúmeras famílias, fléxiveis de distribuições que podem ser usadas em modelos de distribuição desconhecidos.

De entre os métodos de estimação não-paramétricos vamos abordar o empírico e o *kernel*, e dos métodos paramétricos, vamos abordar o modelo binormal e o modelo binormal próprio. O *rating-method* [32]é usado para estimar curvas ROC cujos testes de diagnóstico envolvam um "leitor" e um "intérprete", como por exemplo as imagens radiológicas ou a classificação psiquiátrica. Neste método, os resultados dos testes são classificados em categorias ordenadas que representam o nível de confiança que o "leitor" tem relativamente à presença da doença. Pares de proporção de verdadeiros positivos e de proporção de falsos positivos, resultantes desta classificação, são

a base da estimação da curva ROC.

O método mais comum para a estimação da curva ROC, é o Modelo Binormal. Ajusta uma curva a partir das estimativas de Máxima Verosimilhança para os parâmetros de localização e escala das duas distribuições Gaussianas latentes, subjacentes às populações de indivíduos doentes e não doentes. No entanto com a evolução informática e da metodologia, apareceram novos modelos paramétricos que podem ser usados para estimar curvas ROC. Tosteson e Begg [42] demonstraram que a curva ROC a partir de dados ordinais é ajustada usando modelos de regressão ordinal. Diamond [10] propõe o uso de distribuições logísticas com variâncias iguais.

2.3.1 Curva ROC Empírica

Sejam duas amostras aleatórias $(X_1, X_2, ..., X_m)$ de dimensão m e $(Y_1, Y_2, ..., Y_n)$ de dimensão n, que representam os valores do teste de diagnóstico para os indivíduos não doentes e doentes respectivamente. F_m e G_n são as respectivas funções de distribuição empíricas. Seja c o ponto de corte que estabelece a regra de classificação binária já anteriormente definida. A sensibilidade e a especificidade são estimadas da seguinte forma:

$$\widehat{S} = 1 - G_n(c)$$
$$\widehat{E} = F_m(c)$$

Então a curva ROC empírica é definida por todos os pares (1-$\widehat{E}$, $\widehat{S}$) para todos os valores de c_i que variam na amostra combinada dos valores do teste de diagnóstico para os indivíduos doentes e não doentes. Pontos adjacentes (sem empates na amostra combinada) são ligados por segmentos de recta verticais ou horizontais, que resultam numa curva em escada. As linhas diagonais que possam ocorrer correspondem a empates dos valores do teste de diagnóstico nos dois grupos. Geralmente os pontos de corte c_i não são indicados no gráfico, no entanto cada ponto no gráfico corresponde ao intervalo do tipo $[c_i; c_{i+1})$, cada segmento pode ser associado a um ponto de corte c_i. Se o valor observado corresponder a um indivíduo não doente, o segmento é horizontal se o valor for de um indivíduo doente, o segmento é vertical [5].

Exemplo de Aplicação

Uma vez que a curva ROC é invariante para a tranformação de escala da variável resposta, vamos considerar a tranformação logarítmica dos dados das duas populações definidas na secção 1.2.

A curva ROC empírica para o teste do diagnóstico IgE Total resultante é dada pela figura 2.5.

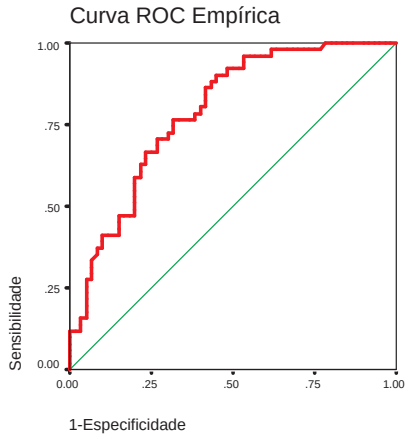

Figura 2.5: Curva ROC empírica

Para construir a curva ROC empírica recorremos ao *package* SPSS (em Anexo Outputs: Curva ROC empírica). Esta aplicação está disponível desde a versão 9.0.

2.3.2 Método de Estimação Kernel

Vamos considerar as amostras e respectivas funções definidas anteriormente. As m observações da variável resposta para os indivíduos não doentes formam o histograma como estimador empírico da função de densidade teórica f_m e de modo semelhante as n observações da variável resposta para os indivíduos doentes formam o histograma como estimador empírico da função de densidade teórica g_n. O método *kernel* emprega técnicas de alisamento às funções de densidade associadas aos valores do teste de diagnóstico dos indivíduos doentes e não doentes, para assim produzir uma curva contínua e suave.

O estimador *kernel* de uma função de densidade de probabilidade é dado por:

$$\widehat{f}(x) = \frac{1}{mh} \sum_{i=1}^{m} K\big(\frac{x - X_i}{h}\big) \tag{2.3}$$

onde K é a função de densidade *kernel*, que satisfaz as condições: $K(.) \geq 0$ e $\int_{-\infty}^{+\infty} K(.)d. = 1$; h é o parâmetro de alisamento, denominado por amplitude (*bandwith*) o qual deve ser fixado pelo utilizador. Este estimador pode ser definido como uma média ponderada dos valores observados em torno de x, onde os pesos são determinados por K [23].

O estimador definido em (2.3) é assintoticamente não enviesado ($E(\widehat{f}(x)) \rightarrow f(x)$) e consistente ($var(\widehat{f}(x)) \rightarrow 0$) sob as seguintes condições: $h \rightarrow 0$ à medida que $m \rightarrow \infty$ e $mh \rightarrow \infty$ à medida que $m \rightarrow \infty$.

Dada a variedade de estimadores *kernel*, Shapiro *et al* [48] utilizam o *kernel biweight*:

$$K(c) = \frac{15}{16}(1 - c^2)^2, \qquad \text{para } c \in [-1, 1] \tag{2.4}$$

Faraggi *et al* [12] utilizam o *kernel* gaussiano:

$$k(c) = (\frac{1}{2\pi})^{1/2} exp(-\frac{c^2}{2}), \qquad \text{para } -\infty < c < +\infty \tag{2.5}$$

A escolha da amplitude h, depende da dimensão da amostra, da função de densidade $f_m(x)$ e da função $K(.)$. Shapiro *et al* [48] utilizam a amplitude:

$$h_m \approx 0.9 min(s_x, \frac{iqr_x}{1.34}) m^{\frac{-1}{5}} \tag{2.6}$$

onde s_x e iqr_x são o desvio-padrão e a amplitude interquartílica empíricos da amostra dos valores do teste de diagnóstico dos indivíduos não doentes. De modo equivalente calcula-se para a amostra de doentes h_n.

A escolha desta amplitude é óptima quando os histogramas têm a forma de sino. Por vezes antes de aplicarmos o método *kernel*, devemos transformar os dados por forma a ficarem simétricos, caso seja necessário (por exemplo a transformação logarítmica ou outro tipo dentro das transformações Box-Cox).

O estimador $\widehat{f}(x)$, na prática substitui cada valor x_i da amostra dos valores do teste de diagnóstico dos indivíduos não doentes, por uma pequena curva com área igual a $1/m$, centrada em x_i, onde cada curva tem a forma de $K(.)$ com amplitude h_m. Depois, unem-se todas as curvas para criar um histograma suave.

A sensibilidade (S) e a especificidade (E) são calculados a partir da versão suavizada $\widehat{F}_m(c)$ e $\widehat{G}_n(c)$:

$$E = \widehat{F}_m(c)$$
$$S = 1 - \widehat{G}_n(c)$$

A curva ROC suave resulta a partir da representação no plano unitário dos pares (1-E, S) para todos os possíveis pontos de corte c.

Exemplo de Aplicação

Pretende-se construir a curva ROC pelo método *kernel* Gaussiano e pelo método *kernel Biweight* para o teste diagnóstico IgE Total, e compará-los entre si e com o método empírico. Vamos utilizar o logarítmo dos dados, uma vez que o método *kernel* terá melhores resultados se os dados forem simétricos.

Método *Kernel* Gaussiano

Foi utilizada uma subrotina do FORTRAN (ver em Anexo Programas: Subrotina *kernel* gaussiana) por forma a construir as funções de densidade pelo método *kernel* Gaussiano (2.5). A partir da equação (2.6) calculámos as amplitudes para cada uma das amostras:

$$h_{nodoentes} = 0.9 * min(1.38; \tfrac{2.0466}{1.34}) 60^{-\frac{1}{5}} = 0.55$$
$$h_{doentes} = 0.9 * min(1.163; \tfrac{1.6127}{1.34}) * 51^{-\frac{1}{5}} = 0.477$$

e para cada uma das amostras obtivémos as seguintes funções de densidade:

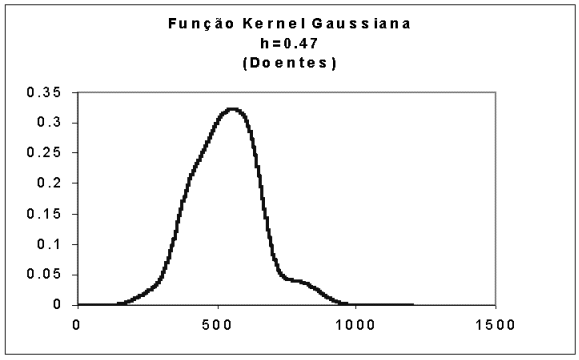

Figura 2.6: Kernel gaussiana para a amostra de doentes

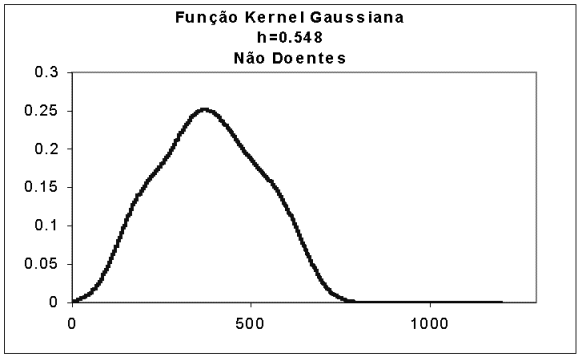

Figura 2.7: Kernel gaussiana para a amostra de não doentes

Podemos verificar que a amplitude proposta pela equação (2.6) não é óptima. Por forma a obter funções de densidade suaves, fomos aumentando h (uma vez que n é fixo) até obtermos f.d.p. suaves.

No gráfico 2.8, podemos observar três densidades que correspondem às amplitudes $h = 0.8$, $h = 0.9$ e $h = 1$.

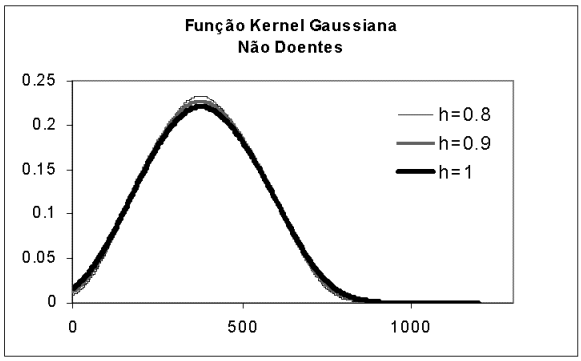

Figura 2.8: Três amplitudes para a amostra de não doentes

Para a amostra de não doentes tomámos $h = 0.8$.

No gráfico 2.9 apresentamos as densidades para as amplitudes $h = 0.9$, $h = 1$ e $h = 1.2$.

Para amostra de doentes tomámos $h = 1$.

Para construir a curva ROC, foi criado um programa em FORTRAN (ver Anexo Programas: Curva ROC *Kernel*), e obtivémos a curva ROC representada na figura 2.10.

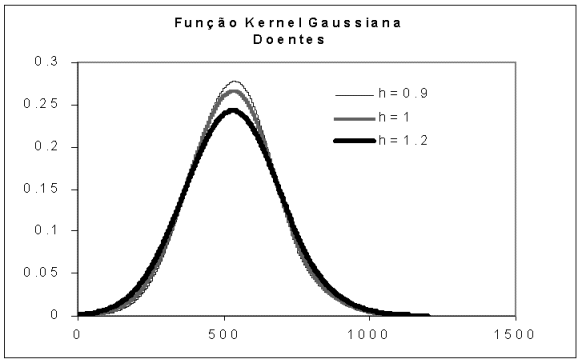

Figura 2.9: Três amplitudes para a amostra de doentes

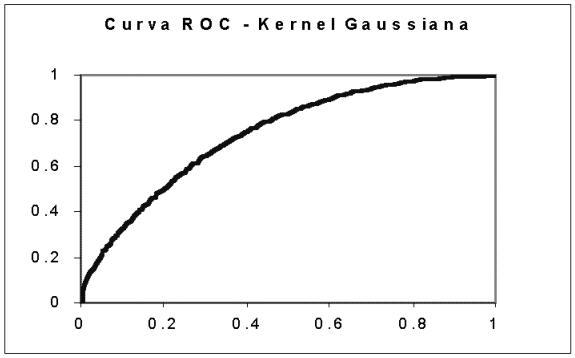

Figura 2.10: Curva ROC *kernel* gaussiana

Podemos observar que a curva ROC resultante é suave.

Método *Kernel* Biweight

Para obtermos as funções de densidade suaves a partir do método *kernel* Biweight (2.4), foi utilizada uma subrotina em FORTRAN (ver Anexo Programas: Subrotina *kernel biweight*). A partir das amplitudes propostas pela equação (2.6) e já anteriormente calculadas, $h_{naodoentes} = 0.55$ e $h_{doentes} = 0.477$ obtivémos as seguintes funções de densidade para as respectivas amostras:

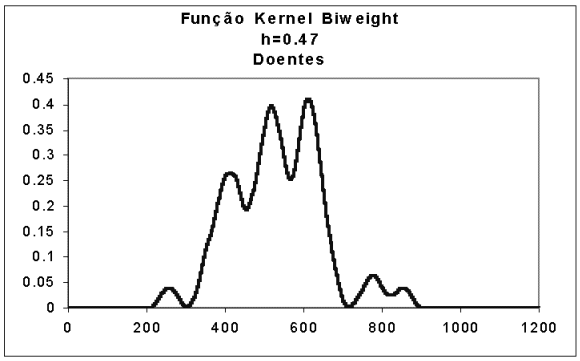

Figura 2.11: Kernel biweight para amostra de doentes

Podemos verificar de que não se tratam de amplitudes óptimas. Como anteriormente já foi feito, vamos aumentado a amplitude até obtermos funções de densidade suaves. Para cada uma das amostras propomos três amplitudes (figuras 2.13 e 2.14)

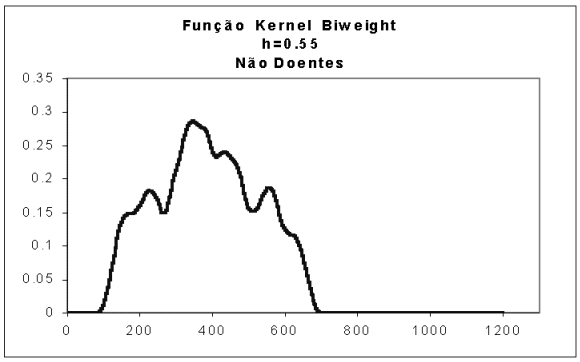

Figura 2.12: kernel biweight para amostra de não doentes

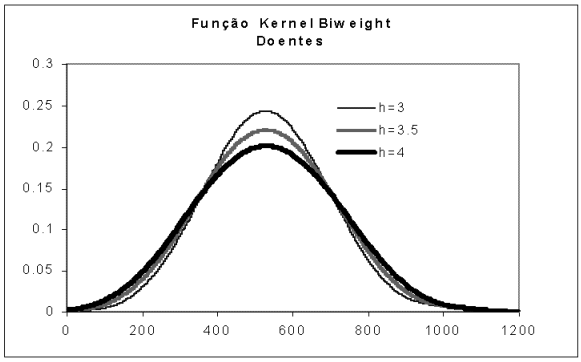

Figura 2.13: Três amplitudes para a amostra de doentes

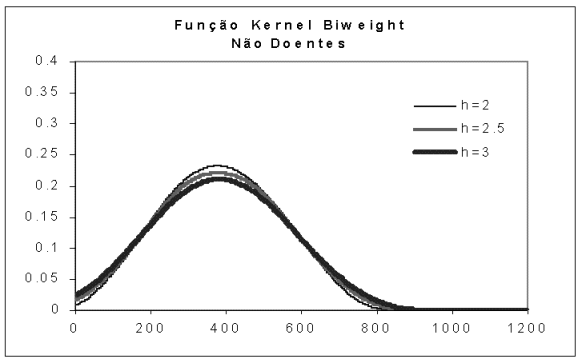

Figura 2.14: Três amplitudes para a amostra de não doentes

Para a amostra de doentes tomámos $h = 3$ e para a amostra de não doentes tomámos $h = 2$.

Obteve-se a seguinte curva ROC:

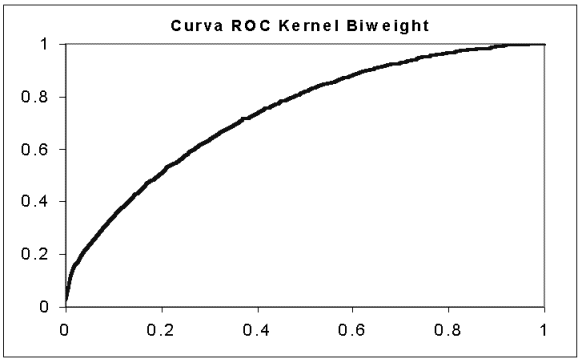

Figura 2.15: Curva ROC *kernel biweight*

Conclusões

A amplitude definida pela expressão (2.6) não se revelou óptima nos dois métodos definidos pelas equações (2.5) e (2.4), embora no método *kernel* Gaussiano a amplitude se aproxime mais da óptima.

Ao comparar as densidades para cada uma das amostras doentes (figura 2.16) e não doentes (figura 2.17), com as amplitudes propostas em cada uma das funções *kernel*, e as respectivas curvas ROC (figura 2.18), verificamos que não apresentam diferenças significativas.

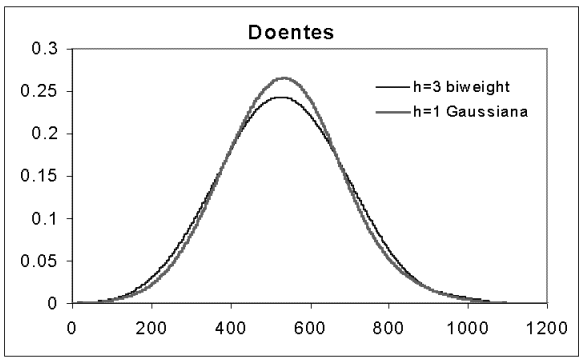

Figura 2.16: Comparação das distribuições de doentes em cada função kernel

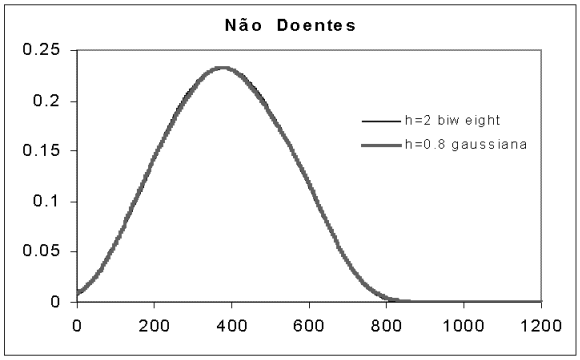

Figura 2.17: Comparação das distribuições de não doentes em cada função kernel

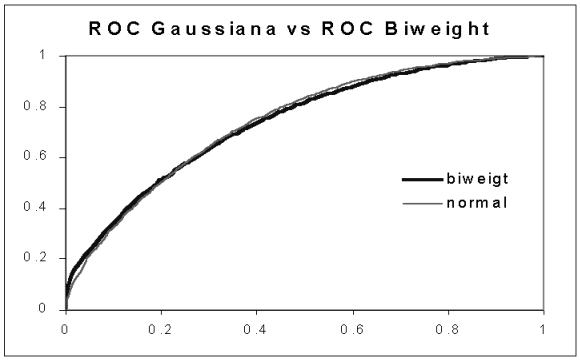

Figura 2.18: Roc Gaussiana vs ROC Biweight

Ao compararmos os dois métodos não-paramétricos, empírico e *kernel*, podemos logo verificar que as respectivas curvas ROC diferem significativamente quanto à suavidade, como seria de esperar. A partir do gráfico 2.19 podemos observar que a curva empírica é mais optimista que a *kernel*, isto é, está mais próxima do vértice (0,1) do plano unitário.

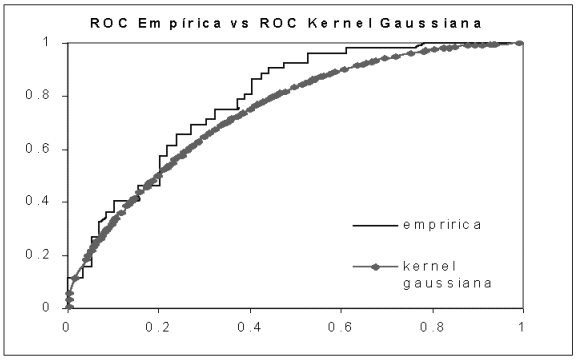

Figura 2.19: ROC empírica vs ROC kernel

2.3.3 Modelo Binormal

Este modelo foi inicialmente criado para testes de diagnóstico cuja variável resposta é ordinal (*rating-method*) [32]. No entanto o método proposto por Metz *et al* [32] ajusta também uma curva ROC a dados contínuos, discretizando primeiramente os dados em categorias, e depois ajustando o modelo binormal a esses dados como se fossem dados ordinais.

O modelo binormal tem como pressuposto a distribuição Normal para a variável resposta dos indivíduos não doentes e dos indivíduos doentes, geralmente com diferentes médias e variâncias.

Se X designar a variável aleatória que representa os valores da variável resposta para os indivíduos não doentes e Y a variável aleatória que representa os valores da variável resposta para os indivíduos doentes, assume-se então que $X \backsim N(\mu_x, \sigma_x^2)$ e $Y \backsim N(\mu_y, \sigma_y^2)$ e independentes. Sem perda de generalidade considera-se que $\mu_x < \mu_y$. Seja c um ponto de corte e admite-se que para valores superiores da variável resposta o indivíduo é classificado de doente e para valores da variável resposta inferiores a este ponto, o indivíduo é classificado de não doente.

Assim sob o modelo binormal, a sensibilidade e a especificidade, para um dado ponto de corte c, são dadas pelas expressões:

$$E(c) = P(X \leq c) = \Phi(\frac{c - \mu_x}{\sigma_x}) \tag{2.7}$$

$$S(c) = 1 - P(Y \leq c) = 1 - \Phi(\frac{c - \mu_y}{\sigma_y}) = \Phi(\frac{\mu_y - c}{\sigma_y}) \tag{2.8}$$

onde $\Phi(.)$ designa a função distribuição Normal reduzida.

Se eventualmente as variáveis X e Y não tiverem uma distribuição Normal, os pressupostos da Binormal postulam a existência de uma transformação ordenada a preservar, seja $h(.)$ a transformação aplicada à variável resposta de tal modo que, $h(X) \backsim N(\mu_x, \sigma_x^2)$ e $h(Y) \backsim N(\mu_y, \sigma_y^2)$, para quaisquer distribuições de X e Y. Uma das propriedades da curva ROC é ser invariante para qualquer transformação de escala monótona à variável resposta (secção 2.2), esta transformação não afecta os resultados.

A curva ROC resulta da representação dos pares $(1 - E, S)$ no plano unitário para todos os pontos de corte c. As equações (2.7) e (2.8) levam a:

$$-c = \sigma_x \Phi^{-1}(1 - E) - \mu_x = \sigma_y \Phi^{-1}(S) - \mu_y$$

que é equivalente a:

$$\frac{\sigma_x}{\sigma_y}\Phi^{-1}(1-E) + (\frac{\mu_y - \mu_x}{\sigma_y}) = \Phi^{-1}(S) \tag{2.9}$$

isto é

$$\Phi^{-1}(S) = a + b\Phi^{-1}(1-E) \tag{2.10}$$

com

$$a = \frac{\mu_y - \mu_x}{\sigma_y}$$

$$b = \frac{\sigma_x}{\sigma_y}$$

A equação (2.10) leva a interpretar a curva ROC binormal como uma recta de declive b e ordenada na origem a, quando representamos nos eixos das abcissas os quantis de probabilidade (1-E) da Normal reduzida (z_{1-E}) e no eixo das ordenadas os quantis de probabilidade S da Normal reduzida (z_S), para todos os pontos de corte c. Neste caso diz-se que representamos a curva no plano binormal (figura 2.20).

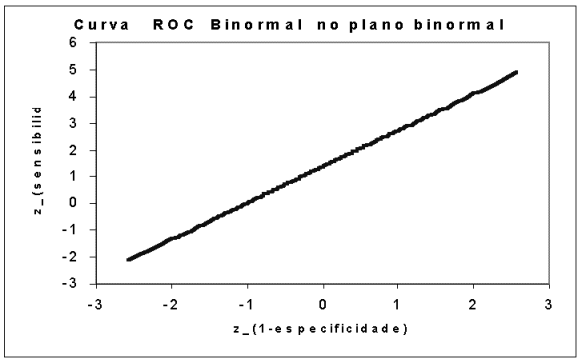

Figura 2.20: Plano Binormal

Função de Verosimilhança

Inicialmente o modelo binormal foi criado para ajustar curvas ROC, onde as variáveis resposta dos testes de diagnóstico eram categóricas em escala ordinal.

O eixo da variável de decisão é particionado em I categorias por $I-1$ pontos de corte, $c_i (i = 1, ..., I-1)$. A probabilidade de resposta na categoria $''i''$ é dada por p_i para os indivíduos não doentes, e q_i para os indivíduos doentes. Os dados das I categorias são obtidos a partir de n indivíduos não doentes e m indivíduos doentes, e o verdadeiro estado de cada indivíduo que produz cada resposta é conhecido pelo investigador. Os dados consistem em $\mathbf{k} = \{k_1, k_2, ..., k_I | \sum_i k_i = m\}$ respostas para os casos não doentes e $\mathbf{l} = \{l_1, l_2, ..., l_I | \sum_i l_i = n\}$ respostas de indivíduos doentes, onde k_i e l_i representam a frequência absoluta de não doentes e doentes respectivamente na categoria i. Estes vectores têm distribuição Multinomial independentes.

Considerando:

- $$c_0 = -\infty$$

  e

  $$c_I = +\infty$$

- F e G funções de distribuição da variável resposta dos indivíduos não doentes e doentes respectivamente.

$$p_i = F(c_i) - F(c_{i-1}) \tag{2.11}$$

$$q_i = G(c_i) - G(c_{i-1}) \tag{2.12}$$

A função de verosimilhança dos parâmetros relativamente aos dados é dada pela expressão:

$$L(\mathbf{k}, \mathbf{l}|a, b, \mathbf{c}) \propto (p_1)^{k_1}(p_2)^{k_2}...(p_I)^{k_I}(q_1)^{l_1}(q_2)^{l_2}...(q_I)^{l_I} \tag{2.13}$$

com as restrições $\sum k_i = m$, $\sum l_i = n$, $\sum p_i = 1$ e $\sum q_i = 1$

Aplicando o logaritmo à função (2.13):

$$lnL \propto \sum_{i=1}^{I}(k_i ln(p_i)) + \sum_{i=1}^{I}(l_i ln(q_i)) \tag{2.14}$$

Ao assumirmos a curva ROC em função de dois parâmetros a e b, a estimativa da curva ROC é obtida a partir de $I + 1$ parâmetros:

$$\Theta = \{a, b, \mathbf{c}\}, \frac{\partial(ln\lambda)}{\partial\theta_j} = 0 \tag{2.15}$$

que conduz a um sistema de $I + 1$ equações não lineares, que pode ser resolvido a partir de vários algoritmos. Os programas "RSCOREII" (Dorfman, 1982) e "ROCFIT" (Metz, 1989) foram desenvolvidos para estimar a

função verosimilhança máxima da curva ROC Binormal a partir do método "*scoring*", o qual é semelhante ao método iterativo de *Newton-Rapson* sendo a matriz Hessiana substituida pela matriz de Informação de Fisher.

Metz *et al* [32] desenvolveram dois algoritmos em FORTRAN, "LABROC4" e "LABROC5" para estimar a função máxima de verosimilhança da curva Binormal para dados com distribuição contínua. Estes dois algoritmos têm por base o algoritmo ROCFIT, que é aplicado a dados categóricos em escala ordinal. Para se aplicar os dois algoritmos, os dados de natureza contínua têm de ser transformados em dados categóricos, particionando o eixo da variável resposta em intervalos adjacentes com fronteiras fixas $w_{i-1} < w_i < w_{i+1}$. Sem perda de generalidade vamos considerar que diferentes intervalos têm diferentes amplitudes. Considerando F e G como definimos anteriormente, e em analogia com as equações (2.11) e (2.12) , a probabilidade de resposta na categoria i de cada indivíduo não doente é $p_i = F(w_i) - F(w_{i-1})$ e $q_i = G(w_i) - G(w_{i-1})$ para cada indivíduo doente. Sejam m indivíduos não doentes e n indivíduos doentes, o logaritmo da função verosimilhança dos dados expresso pela expressão (2.14), onde $\{..., k_{i-1}, k_i, k_{i+1}, ... | \sum_i k_i = m\}$ e $\{..., l_{i-1}, l_i, l_{i+1}, ... | \sum_i l_i = n\}$ representam o número de respostas em cada categoria resultantes dos indivíduos não doentes e doentes respectivamente.

LABROC4 *vs* LABROC5

Os algoritmos LABROC4 e LABROC5 foram desenvolvidos em FORTRAN para a estimação de máxima verosimilhança da curva ROC Binormal para dados com distribuição contínua. O LABROC4 é uma versão do ROCFIT, ajustado ao problema de existir um elevado números de categorias (*truth-state runs*) que podem ocorrer após a categorização dos dados.

O LABROC5 é semelhante ao LABROC4, mas executa o algoritmo ROCFIT só depois de se reduzir o número de categorias por agregação de categorias. O LABROC4 tem uma boa performance, no sentido de produzir boas estimativas até 400 categorias. A partir deste número deve-se utilizar o LABROC5.

Estes dois algoritmos produzem boas estimativas dos parâmetros a e b da curva ROC Binormal, assim como para a área abaixo da curva (AUC), e erros padrão associados a estas estimativas.

Exemplo de aplicação modelo Binormal

Os dados que vamos utilizar referem-se ao logarítmo dos valores do teste de diagnóstico IgE Total descritos no capítulo 1.2.

Para ajustar o modelo binormal aos dados, temos que calcular as estimativas de máxima verosimilhança de a e b. Para calcular as respectivas estimativas, fomos utilizar o programa "ROCKIT 0.9B" [30], e obtiveram-se os seguintes valores (ver Anexo: Output modelo binormal):

$$a = 1.3604 \text{ e } b = 1.3560$$

De seguida apresentam-se as curvas ROC binormais no plano binormal (figura 2.21) e no plano unitário (figura 2.22):

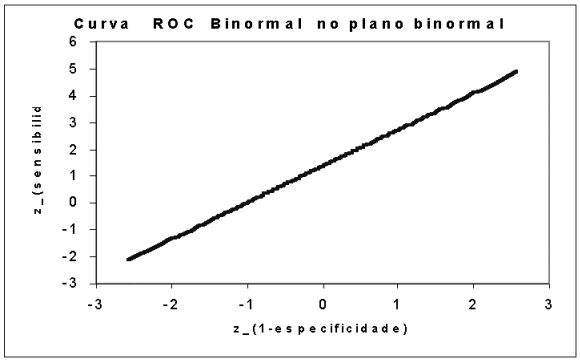

Figura 2.21: Representação da curva ROC Binormal no plano Binormal

Vamos agora comparar a curva ROC Kernel Gaussiana com a curva ROC Binormal no plano unitário (figura 2.23)

E pode observar-se que ambas as curvas são suaves, mas o modelo binormal apresenta-se mais optimista, isto é, mais próximo do vértice (0,1) do plano unitário.

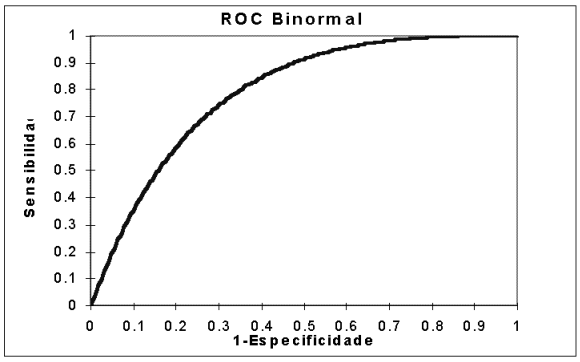

Figura 2.22: Representação da curva ROC Binormal no plano unitário

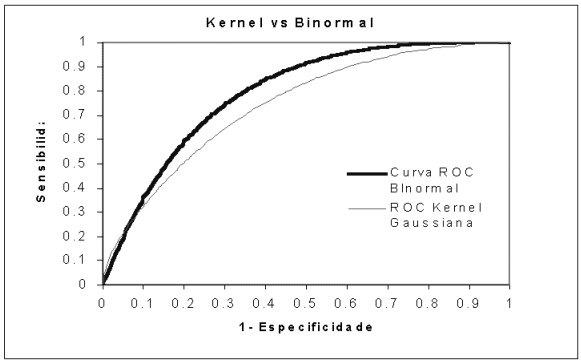

Figura 2.23: ROC Kernel vs ROC Binormal

2.3.4 Modelo Binormal Próprio

A curva ROC binormal convencional utiliza dois parâmetros a e b para especificar a curva ROC. Quando $b \neq 1$, a curva ROC produzida pelo modelo binormal convencional, está sempre acima da diagonal positiva do plano unitário. Quando o modelo binormal convencional é utilizado para ajustar uma curva ROC e a dimensão dos dados é pequena, ou a variável resposta é discreta ordinal existindo uma má distribuição dos dados pelas categorias, pode-se ter porções da curva que caem abaixo da diagonal positiva do plano unitário. Costuma dizer-se que a curva apresenta "ganchos" (figura 2.24). Em situações extremas pode obter-se curvas degeneradas com a forma de "zig-zag" [32].

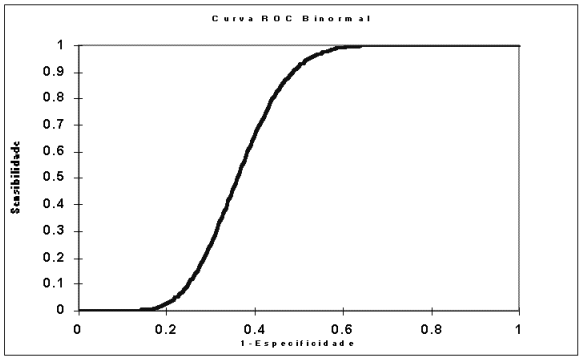

Figura 2.24: Curva ROC degenerada

Metz e Pan [28] desenvolveram um programa em FORTRAN, "PROPROC", que implementa a teoria e algoritmos discutidos em [28], para ajustar uma curva ROC binormal "própria" a dados discretos ordinais e contínuos.

Metz e Pan [28] propõem dois parâmetros para a curva ROC:

$$d'_e = \frac{2a}{b+1} \tag{2.16}$$

e

$$t = \frac{b-1}{b+1} \tag{2.17}$$

Considerando que as duas distribuições subjacentes ao modelo binormal convencional e "próprio" são Gaussianas, o parâmetro d'_e representa o quociente entre a diferença das médias das duas distribuições e a média dos dois desvios padrão e c representa o quociente entre a diferença dos dois desvios padrão e a sua soma.

As curvas ROC binormal convencional e a "própria" são idênticas quando $b \to 1$ e $c \to 0$.

Por forma a simplificar a expressão da área abaixo da curva (AUC), Metz e Pan [28] propuseram outro parâmetro:

$$d_a = \frac{\sqrt{2}a}{\sqrt{1+b^2}} = \frac{d'_e}{\sqrt{1+t^2}} \tag{2.18}$$

d_a representa o quociente entre a diferença entre as duas médias e a raíz quadrada dos dois desvios padrão.

O programa "PROPROC", calcula as estimativas de máxima verosimilhança de d'_a e t.

As curvas ROC binormal convencional e a binormal " própria" são distintas, se a existência de "ganchos" é evidente na curva ROC binormal convencional, empiricamente, isto acontece quando

$$d_a \leq 6 \mid t \mid \tag{2.19}$$

Exemplo de aplicação Modelo Binormal Próprio

Vamos utilizar os dados descritos no capítulo 1.2. Pretende-se ajustar o modelo Binormal "próprio" ao logarítmo dos valores do teste de diagnóstico IgE Total.

Utilizámos o programa "PROPROC" (ver *output* em Anexo: Modelo binormal próprio) para calcular as estimativas de máxima verosimilhança de d_a e t tendo obtido:

$$d_a = 1.0560 \text{ e } t = 0.2247$$

Para construir a curva ROC temos que proceder à conversão dos parâmetros d_a e t, para a e b. De acordo com as equações (2.16) e (2.17) temos:

$$a = d_a \frac{\sqrt{1+b^2}}{\sqrt{2}}$$
$$b = \frac{-(t+1)}{t-1}$$

obtém-se então $a = 1.4$ e $b = 1.58$.
A figura 2.25 representa a curva binormal própria no plano unitário.

Para verificar a existência de ganchos na curva binormal convencional pela expressão (2.19).

Dado que:

$$1.0560 \leq 6 \times 0.2247 = 1.3482$$

podemos concluir que a curva não apresenta "ganchos".

Na figura 2.26 representamos as curvas ROC pelos dois modelos binormais, convencional e próprio:

Através da representação gráfica das duas curvas, podemos observar que parece não existir diferenças significativas entre elas.

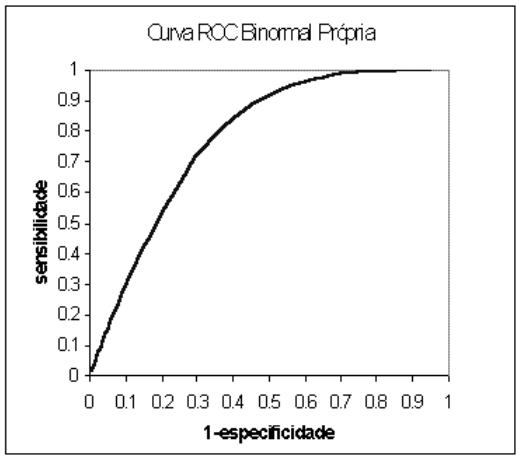

Figura 2.25: Curva ROC binormal própria

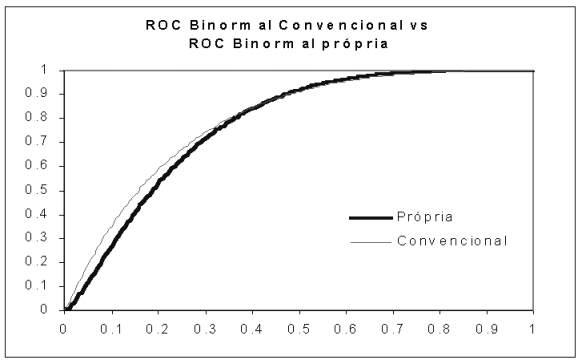

Figura 2.26: Curva ROC binormal convencional *vs* Curva ROC binormal própria

Capítulo 3

Área Abaixo da Curva ROC (AUC)

A área abaixo da curva (AUC) é um dos índices mais utilizados para descrever a exactidão de um teste de diagnóstico. Como de costume vamos admitir que indivíduos doentes apresentam o resultado do teste de diagnóstico mais elevado do que os indivíduos não doentes. A AUC corresponde à probabilidade da variável aleatória que representa os valores da variável resposta dos indivíduos não doentes ser inferior à variável aleatória que representa os valores da variável aleatória dos indivíduos doentes, $AUC = P(X < Y)$ (demonstração na secção 2.2). Esta probabilidade pode ser entendida como a probabilidade do teste classificar correctamente um indivíduo como doente ou não doente. Este índice varia de 0.5 a 1, e quanto mais próximo de 1 melhor é o teste de diagnóstico, isto é, maior é o seu poder de discriminação entre as duas populações. Para classificar os testes de diagnóstico de acordo com a AUC, estabeleceram-se critérios que se apresentam na tabela 3.1.

O cálculo da AUC depende do método pelo qual a curva é estimada. Neste capítulo vamos abordar técnicas não-paramétricas e paramétricas para

AUC	**Teste de Diagnóstico**
0.9 a 1	excelente
0.8 a 0.9	bom
0.7 a 0.8	razoável
0.6 a 0.7	fraco
0.5 a 0.6	mau
< 0.5	!

Tabela 3.1: Valores de referência para classifcar um teste de diagnóstico

estimar a AUC. De entre os métodos não-paramétricos, vamos abordar a estatística de Wilcoxon, a regra do trapézio e o método *kernel*. Relativamente aos métodos paramétricos, vamos abordar o modelo binormal e o modelo binormal próprio.

Estes métodos vão ser aplicados à estimação da AUC para o teste de diagnóstico IgE Total descrito no capítulo 1.2. Relativamente à estimação da AUC pela estatística de Wilcoxon para testes de diagnóstico qualitativos (variável resposta em escala ordinal) utilizámos uma amostra fictícia a título de exemplo.

3.1 Métodos não-paramétricos

3.1.1 Estimação da AUC pela Estatística de Wilcoxon

A estatística de Wilcoxon (W) é usualmente utilizada quando o teste de diagnóstico é qualitativo e a variável resposta tem uma escala ordinal. Este é um método não paramétrico, uma vez que não são exigidos pressupostos distribucionais relativamente às populações de doentes e não doentes. A estatística W não depende dos valores da variável resposta, mas sim das suas ordens.

Seja θ a probabilidade dos valores da variável resposta da população de doentes (Y) ser superior aos valores da variável resposta da população de não doentes (X), $\theta = P(X < Y)$. A estatística W pode ser usada para testar a hipótese: $H_0 : \theta = 0.5$ *vs* $H_1 : \theta > 0.5$, onde a hipótese nula corresponde ao teste de diagnóstico sem poder de discriminação entre as duas populações (a curva ROC coincide com a recta que une os vértices (0,0) e (1,1) do plano unitário).

A estatística de Wilcoxon [17] é dada pela expressão:

$$W = \frac{1}{nm} \sum_{1}^{n} \sum_{1}^{m} S(y, x)$$

onde

$$S(y,x) = \begin{cases} 1 & \text{se } y > x \\ 0.5 & \text{se } y = x \\ 0 & \text{se } y < x \end{cases}$$

e n é a dimensão da amostra de não doentes e m dimensão da amostra de doentes.

Conceptualmente, consiste em fazer todas as nm comparações entre todos os valores das duas amostras. Obviamente que nem todas as comparações são independentes, inclui-las é pura conveniência [17].

Se os dados forem de natureza contínua, o erro padrão de W, $SE(W)$, é dado por [17]:

$$SE(W) = \sqrt{\frac{\theta(1-\theta) + (m-1)(Q_1 - \theta^2) + (n-1)(Q_2 - \theta^2)}{mn}} \tag{3.1}$$

onde

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

$$Q_2 = \frac{2\theta^2}{1+\theta}$$

A estatística W pode ser considerada como uma estimativa de θ, $\widehat{\theta} = W$. No caso da variável resposta ser discreta, W tenderá a subestimar θ, mas a fórmula (3.1) será útil na mesma.

Existe outra forma de estimar θ, que coincide com a regra do trapézio [17], e é adequada para testes de diagnóstico qualitativos.
Considerando as notações:

- n = dimensão da amostra de não doentes
- x_i = categoria i na amostra de não doentes
- n_{x_i} = número de indivíduos não doentes na categoria i
- m = dimensão da amostra de doentes
- y_i = categoria i na amostra de doentes

- m_{y_i} = número de indivíduos doentes na categoria i
- v = número total de categorias na escala da variável resposta

$$W = \frac{\sum_{i=1}^{v}(n_{x_i}\sum_{j>i}(m_{y_j}) + \frac{1}{3}n_{x_i}m_{y_i})}{nm} \tag{3.2}$$

e para calcular o $SE(W)$ de acordo com a fórmula (3.1), as quantidades Q_1 e Q_2 são dadas pelas expressões:

$$Q_1 = \frac{\sum_{i=1}^{v}(n_{x_i}(\sum_{j>i} n_{y_j})^2 + (\sum_{j>i} n_{y_j})n_{y_i} + \frac{1}{3}(n_{y_i})^2)}{n(m)^2} \tag{3.3}$$

$$Q_2 = \frac{\sum_{i=1}^{v}(n_{y_i}(\sum_{j<i} n_{x_j})^2 + (\sum_{j<i} n_{x_j})n_{x_i} + \frac{1}{3}(n_{x_i})^2)}{m(n)^2} \tag{3.4}$$

Por outro lado, o cálculo da estatística de *Mann-Whitney-Wilcoxon* quando $m > 20$ [35, 9] permite-nos testar $H_0 : \theta = 0.5$ *vs* $H_1 : \theta > 0.5$:
já que, sob H_0

$$Z_0 = \frac{W - \frac{nm}{2}}{\sqrt{\frac{nm(n+m+1)}{12}}} \backsim N(0,1)$$

Para calcular a AUC pela estatística de Wilcoxon e o respectivo erro padrão, o programa "SPSS" calcula esta estatística e respectivo erro padrão para dados de natureza contínua e discreta e o programa "ROCKIT" calcula-os apenas para dados de natureza contínua.

3.1.2 Estimação da AUC pela regra do trapézio

A regra do trapézio é um método numérico de integração para estimar a área abaixo de uma curva. O intervalo de integração [a,b] é dividido em n sub-intervalos. Em cada sub-intervalo a função é aproximada por uma função linear. O integral da função linear em cada sub-intervalo corresponde à área de um trapézio. Somando essas áreas produz a desejada aproximação do integral definido. Quanto maior for n melhor será a aproximação para o valor exacto do integral.

Método:

Dada uma função f e os limites do intervalo a e b, definimos as amplitudes dos sub-intervalos por $h = \frac{b-a}{n}$ e os pontos da subdivisão por $x_k = a + kh, k = 0, ..., n$, a expressão que a seguir se apresenta corresponde ao valor aproximado da área abaixo da curva:

$$T_n = h \times (\frac{f(x_0)}{2} + \sum_{k=1}^{n-1} f(x_k) + \frac{f(x_n)}{2}) \tag{3.5}$$

Assumindo que a segunda derivada da função integranda é contínua no intervalo de integração, temos que:

$$\int_a^b f(x)dx = T_n - \frac{b-a}{12} f''(c)h^2, c \in (a, b) \tag{3.6}$$

Definimos o erro por:

$$E_{T_n} = \int_a^b f(x)dx - T_n$$

e obtemos uma estimativa do erro dada por:

$$\mid E_{T_n} \mid = \frac{b-a}{12} \mid f^{''}(c) \mid h^2 \leq M\frac{b-a}{12}h^2$$

onde M é o limite superior da segunda derivada:

$$\mid f^{''}(x) \mid \leq M$$

para

$$a < x < b$$

Para a aplicação da regra do trapézio, os subintervalos têm de ser igualmente espaçados e ter o conhecimento de $f(x)$. Quando se aplica a regra do trapézio para calcular a AUC da curva ROC, os sub-intervalos (definidos pelos pontos de corte) não são igualmente espaçados e no caso não-paramétrico não temos o conhecimento da função $f(x)$. Então para calcularmos a AUC pela regra do trapézio para a curva ROC, temos de considerar o seguinte: os pontos de corte que definem os valores da sensibilidade e 1-especificidade (coordenadas da curva no plano unitário), vão determinar os sub-intervalos, a área vai corresponder à soma de triângulos e rectângulos por eles definidos:

$$AUC = \frac{((1-E_{i+1})-(1-E_i))(S_{i+1}-S_i)+2((1-E_{i+1})-(1-E_i))(S_i)}{2}$$

para $i = 1...r$ (onde r é o número total de pontos de corte).

Para calcular a AUC pela regra do trapézio, pode-se implementar o algoritmo numa folha de cálculo, por exemplo o "EXCEL".

3.1.3 Estimação da AUC pelo método *kernel*

Se considerarmos o modelo *kernel* Gaussiano (2.5) , a estimativa da AUC é dada por [7]:

$$A_K = \frac{1}{nm} \sum_{i}^{m} \sum_{j=1}^{n} \Phi(\frac{x_i - y_j}{\sqrt{h_x^2 + h_y^2}})$$

Para calcular A_K pode-se implementar o algoritmo numa folha de cálculo, por exemplo no "EXCEL".

3.2 Métodos paramétricos

3.2.1 Estimação da AUC pelo modelo binormal

Este é um método paramétrico, uma vez que são exigidos pressupostos distribucionais subjacentes às populações de doentes e não doentes. Assumindo que X e Y são independentes, $X \sim N(\mu_x, \sigma_x^2)$ e $Y \sim N(\mu_y, \sigma_y^2)$. A especificidade (E) e a sensibilidade (S), para um dado ponto de corte são dadas por:

$$E(c) = P(X < c) = \Phi(\frac{c - \mu_x}{\sigma_x})$$

$$S(c) = P(Y > c) = \Phi(\frac{\mu_y - c}{\sigma_y})$$

A AUC é dada pela expressão [12]:

$$A_z = \Phi(\frac{\mu_y - \mu_x}{\sqrt{\sigma_x^2 + \sigma_y^2}}) \tag{3.7}$$

Com efeito a área abaixo da curva ROC corresponde à $P(X < Y)$, então:

$$P(X < Y) = P(X - Y < 0) = P(\frac{X - Y - (\mu_x - \mu_y)}{\sqrt{\sigma_x^2 + \sigma_y^2}} < \frac{-\mu_x + \mu_y}{\sqrt{\sigma_x^2 + \sigma_y^2}}) = \Phi(\frac{\mu_y - \mu_x}{\sqrt{\sigma_x^2 + \sigma_y^2}})$$

◁

Sabendo que os parâmetros da curva são dados em função de a e b a AUC pode ser dada em função dos parâmetros da curva:

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

Os programas "ROCKIT" e "ROCFIT" calculam as estimativas para A_z e para $SE(A_z)$, para dados de natureza contínua ou discreta.

3.2.2 Estimação da AUC pelo método binormal próprio

Considerando a parametrização proposta por Metz e Pan [28] para a curva ROC:

$$d'_e = \frac{2a}{b+1}$$

e

$$t = \frac{b-1}{b+1}$$

a estimação da AUC para o modelo binormal próprio é dada por [28]:

$$A_{prop} = \Phi(\frac{d'_e}{\sqrt{2(1+t^2)}}) + 2F(\frac{-d'_e}{\sqrt{2(1+t^2)}}, 0; -\frac{1-t^2}{1+t^2}) \tag{3.8}$$

onde $F(X, Y, \rho)$ é a distribuição Gaussiana bivariada padrão com coeficiente de correlação ρ.
Como já foi referido anteriormente a curva binormal convencional e a própria são idênticas quando $b \to 1$ e $t \to 0$, assim o segundo termo da expressão (3.8) desaparece, e o primeiro termo é equivalente à AUC da curva binormal convencional A_z.

Considerando a reparametrização no modelo binormal próprio $d_a = \frac{d'_e}{\sqrt{1+t^2}}$ a AUC é dada por:

$$A_{prop} = \Phi(\frac{d_a}{\sqrt{2}}) + 2F(-\frac{d_a}{\sqrt{2}}, 0; -\frac{1-t^2}{1+t^2}) \tag{3.9}$$

Esta nova parametrização na curva binormal própria torna o cálculo da AUC mais simples. Se $t = 0$ (i.e. $b = 1$) o segundo termo da expressão (3.9)

é igual a zero, e o primeiro termo é equivalente a A_z. O segundo termo é positivo se $t \neq 0$, indicando que a A_{prop} é superior a A_z.

O programa "PROROC" calcula a estimativa para A_{prop}, mas não calcula os erros padrão das estimativas dos parâmetros da curva ROC nem a estimativa do erro padrão da área.

3.3 Exemplos de aplicação

O objectivo nesta secção é de exemplificar a aplicação da estatística de Wilcoxon para o cálculo da AUC para um teste de diagnóstico qualitativo e comparar os vários métodos para o cálculo da AUC aqui propostos e os respectivos erros padrão.

3.3.1 Estatística de Wilcoxon para um teste diagnóstico qualitativo

Vamos considerar um teste de diagnóstico cuja variável resposta assume cinco valores: 1,2,3,4,5 e vamos considerar a amostra da tabela 3.2

Considerando a expressão (3.2) para o cálculo da AUC e para o cálculo do erro padrão (3.1) temos que calcular as quantidades (3.3) e (3.4). O quadro 3.3 contém os cálculos intermédios:

$$w = \frac{2275}{50 \times 50} = 0.91$$

	1	**2**	**3**	**4**	**5**	**Total**
Doentes	2	4	10	14	20	50
Não Doentes	28	14	5	2	1	50

Tabela 3.2: Amostra (hipotética) de um teste de diagnóstico qualitativo com cinco pontos

	1	**2**	**3**	**4**	**5**	**Total**
N.º Doentes classificados em i	2	4	10	14	20	50
N.º Não Doentes classificados em i	28	14	5	2	1	50
Doentes >i	48	44	34	20	0	
Parcela i de (3.2)	1372	644	195	54	10	
Não doentes <i	0	28	42	47	49	
Parcela i de (3.3)	67237.3	29642.67	7646.67	1490.67	133.33	106150.64
Parcela i de (3.4)	522.67	4965.3	19823.3	32260.67	49006.67	106578.61

Tabela 3.3: Cálculos auxiliares para a estatística de Wilcoxon e Q_1 e para Q_2

$$Q_1 = 0.849$$

e

$$Q_2 = 0.853$$

e o erro padrão é

$$SE(W) = \sqrt{\frac{2.3261}{50 \times 50}} = 0.031$$

3.3.2 Comparação dos vários métodos de estimação da AUC para um teste de diagnóstico quantitativo

Vamos comparar as AUC para cada um dos métodos anteriores e os respectivos erros padrão. O cálculo da estatística de Wilcoxon foi feito usando o *package* estatístico "SPSS", (o programa "ROCFIT" também calcula esta estatística). Para o cálculo da área usando o método *kernel* e a regra do trapézio, implementaram-se os algoritmos em "Excel". O cálculo da AUC pelo modelo binormal convencional foi feito usando o programa "ROCFIT" (ver Anexo *Outputs*: Modelo binormal), e pelo modelo binormal próprio foi feito pelo programa "PROPROC" (ver Anexo *Outputs*: Modelo binormal próprio).

Apresentam-se na tabela 3.4 os valores das AUCs e os respectivos erros padrão.

Conclusão:

A estatística de Wilcoxon e o modelo binormal convencional são os métodos que apresentam AUCs que não diferem significativamente, e comparando-os com o modelo binormal próprio também não parece que a diferença seja significativa. O método *kernel* foi o menos optimista, ou seja foi o que apresentou um valor da AUC mais baixa.

Método	**AUC**	**SE(AUC)**
Wilcoxon	$W = 0.7857$	0.0442
Kernel	$A_K = 0.736$	
Binormal	$A_z = 0.7903$	0.0426
Binormal próprio	$A_{prop} = 0.7724$	

Tabela 3.4: Comparação das AUCs

Ao aplicarmos a regra do trapézio às curvas construidas pelos métodos *kernel*, modelo binormal e modelo binormal próprio (não se aplicou à curva empírica uma vez que esta é irregular), verifica-se que a regra do trapézio coincidiu exactamente com os modelos binormal e binormal próprio, e para o método *kernel* deu um valor de 0.736, comparando com o valor pelo método *kernel*, 0.721, parece não existir diferenças significativas.

Capítulo 4

Comparação de Testes de Diagnóstico

Quando se pretende avaliar a performance de um teste de diagnóstico, construimos a respectiva curva ROC e classificámo-lo de acordo com o valor da AUC (tabela 3.1). Para aplicar um teste de diagnóstico a indivíduos futuros é necessário estabelecer um ponto de corte à priori, mas para a construção da respectiva curva ROC essa decisão não é necessária, tendo em conta que a contrução da curva ROC é obtida através da representação da sensibilidade e 1-especificidade para todos os pontos de corte possíveis no plano unitário. Para comparar vários testes de diagnóstico, independentemente do ponto de corte utilizado em cada um, podemos usar as AUCs.

A comparação de testes de diagnóstico torna-se cada vez mais pertinente, uma vez que a evolução científica e tecnológica contribui para o aparecimento cada vez mais acelerado de novos testes de diagnóstico.
A comparação de dois testes de diagnóstico pode ser feita comparando as respectivas áreas abaixo da curva ROC. Isto o que corresponde a testar a hipótese nula $H_0 : \theta_1 = \theta_2$ onde $\theta_i = P_i(X < Y), i = 1, 2$, onde i representa o teste de diagnóstico, ou então comparar um teste conhecido contra um novo

$H_0 : \theta_1 > \theta_2$ vs $H_0 : \theta_1 < \theta_2$. Essa comparação pode ser feita numa perspectiva paramétrica ou numa perspectiva não-paramétrica.

Neste trabalho vamos abordar seis situações para a comparação das AUC: amostras independentes, amostras emparelhadas e amostras parcialmente emparelhadas, e em cada tipo de amostras a abordagem paramétrica e a abordagem não-paramétrica.

Para amostras independentes Metz [29] propõe um método paramétrico para comparar as AUCs enquanto que Hanley *et al* [18] propõem um método não-paramétrico. Para amostras emparelhadas DeLong *et al* [11] e Hanley *et al* [18] propõem um método não-paramétrico para comparar AUC enquanto que Metz *et al* [31] propõem um método paramétrico. Para amostras parcialmente emparelhadas Zhou *et al* [47] propõem um método não-paramétrico para comparar AUC enquanto que Metz *et al* [31] propõem um método paramétrico.

Para além das comparações das respectivas AUC dos testes de diagnóstico, isto é, uma comparação global da performance dos testes, Bloch [4] propõe um método onde o investigador pode comparar testes sob um critério específico. O que se pretende com este método é comparar os riscos associados a falsos negativos e falsos positivos de cada teste de diagnóstico, o que tiver menor risco é o preferido. Por vezes acontece que os testes de diagnóstico têm áreas (AUC) que não são significativamente diferentes e no entanto podem ter riscos diferentes.

4.1 Comparação de Áreas

4.1.1 Amostras Independentes

Método Paramétrico

Os testes paramétricos para testar a diferença entre as áreas de duas curvas com amostras independentes, baseiam-se nos seguintes pressupostos:

- os indivíduos são seleccionados aleatoriamente a partir de uma população de interesse.
- os indivíduos são seleccionados aleatoriamente para cada um dos testes de diagnóstico.
- os testes de diagnóstico são interpretados independentemente.
- cada indivíduo disponibiliza resultados só para um teste de diagnóstico.
- os dados provêm de um par de distribuições normais geralmente com diferentes médias e variâncias.

Sob a hipótese nula $H_0 : \theta_1 = \theta_2$, de que não existem diferenças significativas entre as AUCs para duas curvas ROC independentes, θ_1 corresponde à AUC para o teste de diagnóstico 1 e θ_2 corresponde à AUC para o teste de diagnóstico. A estatística de teste é dada por [32]:

$$Z = \frac{A_{z1} - A_{z2}}{\sqrt{\widehat{var}(A_{z1}) + \widehat{var}(A_{z2})}} \tag{4.1}$$

onde A_{z1} e A_{z2} são as AUC para cada um dos testes estimadas pelo modelo binormal (3.7).

A_{z1} e A_{z2} são estimadores de máxima verosimilhança, e pelas propriedades destes estimadores têm distribuição assintoticamente Normal padrão, então a diferença também tem distribuição assintoticamente Normal, pelas propriedades da distribuição Normal. Assim, esta estatística de teste, sob hipótese nula, tem uma distribuição assintoticamente Normal padrão.

O programa "ROCKIT" calcula estas estimativas e o valor da estatística de teste para dados de natureza contínua ou discreta.

Método Não-Paramétrico

Hanley *et al* [18] propõem um método não-paramétrico para testar $H_0 : \theta_1 = \theta_2$, θ_1 corresponde à AUC para o teste de diagnóstico 1 e θ_2 corresponde à AUC para o teste de diagnóstico 2, em amostras independentes. A partir da estatística de Wilcoxon define a estatística de teste para comparar áreas da seguinte forma:

$$Z = \frac{A_1 - A_2}{\sqrt{SE_1^2 + SE_2^2}} \tag{4.2}$$

onde A_1 e A_2 são as AUC estimadas para cada um dos testes de diagnóstico a partir da estatística de Wilcoxon e SE_1 e SE_2 são os respectivos erros padrão calculados a partir da expressão (3.1). Esta estatística segue aproximadamente, sob a hipótese nula, uma distribuição Gaussiana padrão.

4.1.2 Amostras Emparelhadas

Método Paramétrico

Metz *et al* [31] propõem um algoritmo paramétrico para testar AUCs de curvas ROC para amostras emparelhadas. Elaboraram um programa em FORTRAN "CORROC", baseado na maximização da função de verosimilhança.

Este algoritmo baseia-se nos seguintes pressupostos:

- a amostra foi recolhida aleatoriamente de uma população de interesse.
- todos os indivíduos foram sujeitos aos dois testes de diagnóstico.
- os resultados dos testes são interpretados de forma independente.
- a variável resposta provem de um par de dsitribuições normais bivariadas geralmente com diferentes médias, variâncias e correlações.

Sob a hipótese nula $H_0 : \theta_1 = \theta_2$, a estatística de teste para a diferença de áreas é dada por [31]:

$$Z_\Delta = \frac{A_{z1} - A_{z2}}{\sqrt{\widehat{var}(\Delta_z)}} \tag{4.3}$$

onde $\widehat{var}(\Delta_z) = var(A_{z1}) + var(A_{z2}) - 2cov(A_{z1}, A_{z2})$

A_{z1} e A_{z2} são estimadores de máxima verosimilhança, então esta estatística de teste, sob a hipótese nula, segue assintoticamente uma distribuição Normal padrão.

O programa "CORROC" calcula o valor da estatística de teste (4.3) para a diferença das AUC para amostras emparelhadas para dados de natureza

discreta ou contínua.

Método Não-Paramétrico

Vamos considerar dois algoritmos, um proposto por Hanley *et al* [18] e outro por DeLong *et al* [44].

Hanley *et al*

Hanley *et al* [18] abordam a situação de duas curvas construidas com base na mesma amostra de indivíduos. Considerando a estatística (4.2) e tendo em consideração que as amostras são emparelhadas a estatística de teste para a hipótese $H_0 : \theta_1 = \theta_2$ é dada por [18]:

$$Z = \frac{A_1 - A_2}{\sqrt{SE_1^2 + SE_2^2 - 2rSE_1SE_2}} \tag{4.4}$$

onde r representa a estimativa da correlação entre A_1 e A_2. O cálculo de r envolve dois coeficientes de correlação intermédios: $r_{naodoentes}$ que representa o coeficiente de correlação da população de não doentes entre os dois testes de diagnóstico e $r_{doentes}$, que representa o coeficiente de correlação da população de doentes entre os dois testes. Estes dois coeficientes podem ser calculados pelo coeficente de correlação de Pearson ou pelo τ de Kendall de acordo com o tipo de dados. Para calcular r, Hanley *et al* disponibilizam uma tabela de dupla entrada (vem em Anexo Tabelas: Tabela de coeficientes de correlação entre áreas), onde em primeiro lugar se tem de calcular a média dos coeficientes de correlação $\frac{r_{naodoentes}+r_{doentes}}{2}$ e a média das áreas $\frac{A_1+A_2}{2}$.

Quanto maior for r mais sensível o teste Z será [18].

Esta estatística de teste, sob a hipótese nula, segue aproximadamente uma distribuição Gaussiana padrão.

DeLong *et al*

DeLong *et al* [44] propõem um método não paramétrico para comparar testes de diagnóstico em amostras emparelhadas, baseado na estatística de Wilcoxon.

Vamos supor que temos $N = m + n$ indivíduos e a variável resposta é contínua e para valores elevados da variável de decisão implica um teste positivo. Vamos também considerar que m indivíduos são doentes e n são não doentes. $x_j, j = 1...m$ representa os resultados do teste de diagnóstico para os indivíduos doentes e $y_k, k = 1...n$ os resultados do teste de diagnóstico para os indivíduos não doentes.

Seja

$$S(x_j, y_k) = \begin{cases} 1 & \text{se } x_j > y_k \\ 0.5 & \text{se } x_j = y_k \\ 0 & \text{se } x_j < y_k \end{cases}$$

então $V(x_j)$ e $V(y_k)$ são definidos por:

$$V(x_j) = \frac{\sum_{i=1}^{n} S(x_j, y_i)}{n}, j = 1...m$$

$$V(y_k) = \frac{\sum_{i=1}^{m} S(x_i, y_k)}{m}, k = 1...n$$

$V(x_j)$ representa a fracção de y *scores* que são inferiores a ele e $V(y_k)$ representa a fracção de x *scores* que são superiores a ele.

De acordo com a estatística de Wilcoxon vem que a AUC é estimada por:

$$\widehat{\theta} = \frac{\sum_{j=1}^{m} V(x_j)}{m} = \frac{\sum_{i=1}^{n} V(y_k)}{n} \tag{4.5}$$

e a variância da AUC pelo método de DeLong é dada por:

$$var(\widehat{\theta}) = \frac{m \times \sum_{j=1}^{m} V(x_j)^2 - (\sum_{j=1}^{m} V(x_j))^2}{m^2(m-1)} + \frac{n \times \sum_{k=1}^{n} V(y_k)^2 - (\sum_{k=1}^{n} V(y_k))^2}{n^2(n-1)} \tag{4.6}$$

A covariância entre as áreas estimadas pelos dois testes de diagnóstico, 1 e 2, é dada por:

$$cov(\widehat{\theta}_1, \widehat{\theta}_2) = \frac{\sum_{j_1}^{m}(V(x_{1j} - \bar{V}(x_1))(V(x_{2j}) - \bar{V}(x_2))}{m(m-1)} + \frac{\sum_{k=1}^{n}(V(y_{1k}) - \bar{V}(y_1))(V(y_{2k}) - \bar{V}(y_2))}{n(n-1)} \tag{4.7}$$

onde $\bar{V}(x)$ e $\bar{V}(y)$ representam a média de $V(x_j)$ e de $V(y_k)$ respectivamente.

Para testar a hipótese $H_0 : \theta_1 = \theta_2$ calcula-se o quociente crítico (CR):

$$CR = \frac{\widehat{\theta}_1 - \widehat{\theta}_2}{SE(\widehat{\theta}_1 - \widehat{\theta}_2)} \tag{4.8}$$

onde $SE(\widehat{\theta}_1 - \widehat{\theta}_2) = \sqrt{var(\widehat{\theta}_1) + var(\widehat{\theta}_2) - 2cov(\widehat{\theta}_1, \widehat{\theta}_2)}$ e sob a hipóetese nula, CR é uma variável aleatória com distribuição aproximadamente Gaussiana padrão.

4.1.3 Amostras Parcialmente Emparelhadas

Nem sempre é possivel obter amostras emparelhadas completas. Existem várias razões geralmente relacionadas com factores indirectamente ligados à investigação, por exemplo, perda de registos, indivíduos que desistem ou morrem, etc. Então podemos ter uma situação em que temos indivíduos que realizaram apenas um dos testes de diagnóstico e outros indivíduos que realizaram os dois testes de diagnóstico. Quando temos as duas situações em simultâneo dizemos que temos amostras parcialmente emparelhadas (tabela 4.1).

Os indivíduos que não realizaram os dois testes de diagnóstico teriam de ser omitidos do estudo, que se traduziria num custo considerável e numa perda de potência estatística.

	Teste 1	**Teste 2**	
Indivíduo 1	X_1		u indivíduos
.	.		não emparelhados
.	.		no teste 1
.	.		
Indivíduo u	X_u		
Individuo $u+1$	X_{u+1}	Y_{u+1}	v indivíduos
.	.	.	emparelhados
.	.	.	nos testes 1 e 2
.	.	.	
Indivíduo$u+v$	X_{u+v}	Y_{u+v}	
Indivíduo $u+v+1$		Y_{u+v+1}	w indivíduos
.		.	não emparelhados
.		.	no teste 2
.		.	
Indivíduo $u+v+w$		Y_{u+v+w}	

Tabela 4.1: Amostras parcialmente emparelhadas

Método Paramétrico

Metz *et al* [31] propõem um método para comparar as AUC para amostras parcialmente emparelhadas, seguindo as metodologias definidas nas secções 4.1.1 e 4.1.2.

O programa "ROCKIT" desenvolvido por Metz [30] implementa este algoritmo para dados de natureza discreta ou contínua.

Método Não-paramétrico

Zhou *et al* [47] desenvolveram um método para comparar curvas ROC para dados parcialmente emparelhados quer de natureza contínua quer de natureza discreta. Construíram um estimador consistente da matriz de covariâncias, baseadas na derivação da matriz de covariâncias dos estimadores não paramétricos das AUC para dados parcialmente emparelhados e utilizam este estimador para testar as diferenças das áreas.

4.2 Comparação de Riscos

4.2.1 Um método proposto por Bloch

Cada ponto da curva ROC descreve a performance do teste de diagnóstico para um dado ponto de corte. Bloch [4] propõe um método para comparar testes de diagnóstico tendo em conta os custos associados a falsos positivos e a falsos negativos. Duas curvas ROC podem ter as correspondentes AUC semelhantes e no entanto terem riscos diferentes, o teste escolhido deverá ser aquele que apresenta risco menor. Portanto, para além de comparar dois testes de uma forma global (comparando as AUC), podemos comparar testes sob uma circunstância específica (comparação de riscos).

Bloch admite que as amostras são emparelhadas, isto é, todos os indivíduos realizaram os dois testes de diagnóstico e o mesmo "gold standard". Diferentes pontos de corte são especificados para cada teste. Considera dois tipos de amostragem: a global, onde uma amostra de dimensão $n = n_1 + n_2$ é seleccionada aleatoriamente de uma população de interesse, e n_1 representa o n.º de indivíduos com a doença e n_2 o n.º de indivíduos sem a doença; e o método de amostragem controlada, onde n_1 indivíduos são seleccionados aleatoriamente de uma população com a doença e n_2 individuos são seleccionados aleatoriamente de uma população sem a doença. Neste trabalho vamo-nos referir à amostragem global.

Riscos para comparar testes de diagnóstico

O tipo de amostragem global permite-nos estimar a sensibilidade a especificidade e a prevalência. Na tabela 4.2 estão definidas as probabilidades conjuntas para os dois testes.

p_{ijk} representa a probabilidade conjunta da performance dos dois testes de diagnóstico, para $i = 1, 2$ (resultado positivo=1 e negativo=2) no teste 1; $j = 1, 2$ resultado no teste 2 e $k = 1, 2$ resultado do "gold standard".

Vamos considerar as seguintes notações:

- $p_1 = P$(teste 1 ser positivo)
- $p_2 = P$(Teste 2 ser positivo)
- $\pi = P$(gold standard positivo) =Prevalência
- L' custo associado a um falso positivo
- L custo associado a um falso negativo
- R_i risco associado ao teste $i = 1, 2$
- E_i especificidade associada ao teste i
- S_i sensibilidade associada ao teste i
- l_{ijk} frequências observadas (ex: l_{111} n.º de indivíduos com os dois testes positivos e a doença está presente, $l_{1..}$ n.º de indivíduos com o teste 1 positivo, $l_{2.1}$ n.º de indivíduos com o teste 1 negativo e o "gold standard" positivo)

	Teste 1 Positivo (1)	**Teste 1 Negativo (2)**	
Teste 2 Positivo (1)	p_{111}	p_{211}	p_2
	p_{112}	p_{212}	
Teste 2 Negativo (2)	p_{121}	p_{221}	$1 - p_2$
	p_{122}	p_{222}	
	p_1	$1 - p_1$	1

Tabela 4.2: Probabilidades conjuntas para os dois testes de diagnóstico

O risco associado a cada teste é dado por:

$$R_i = L'(1-\pi)(1-E_i) + L\pi(1-S_i) \tag{4.9}$$

O risco depende do ponto de corte associado a cada teste e é este que determina a positividade do teste. Uma vez que são estabelecidos pontos de corte diferentes para cada teste é natural que os riscos sejam diferentes.

A tabela 4.2 pode ser expressa em termos de riscos, da prevalência e na probabilidade do resultado do teste ser positivo, então podemos definir o risco por:

$$R_1 = L'(p_1 - p_{1.1}) + L(\pi - p_{1.1}) \tag{4.10}$$

Demonstração:

Por 4.9 vem que $R_1 = L'(1-\pi)(1-E_1) + L\pi(1-S_1)$, então temos que demonstrar que:

(i)$(1-\pi)(1-E_1) = p_1 - p_{1.1}$ e (ii)$\pi(1-S_1) = (\pi - p_{1.1}$

Desenvolvendo a equação (i) em ordem a E_1 vem:

$E_1 = \frac{1-\pi-p_1+p_{1.1}}{1-\pi}$

e pela definição de especificidade, sabemos que:

E_1 =P(teste 1 negativo|g.s. negativo)=$\frac{P(teste1negativo,g.s.negativo)}{P(g.s.negativo)} = \frac{p_{2.2}}{1-\pi}$

Então, tem de se verificar a igualdade $\frac{1-\pi-p_1+p_{1.1}}{1-\pi} = \frac{p_{2.2}}{1-\pi}$, e é suficiente provar que $1-\pi-p_1+p_{1.1} = p_{2.2}$:

$1-\pi-p_1+p_{1.1} = (p_{111}+p_{211}+p_{112}+p_{212}+p_{121}+p_{221}+p_{122}+p_{222})-$
$-(p_{111}+p_{211}+p_{121}+p_{221}) - (p_{111}+p_{112}+p_{121}+p_{122}) = p_{212}+p_{222} = p_{2.2}$

Desenvolvendo a equação (ii) em ordem a S_1 vem que:

$S_1 = \frac{p_{1.1}}{\pi}$

e por definição de sensibilidade:

S_1 =P(teste 1 positivo|g.s. positivo)= $\frac{P(teste1positivo,g.s.positivo)}{P(g.s.positivo)} = \frac{p_{1.1}}{\pi}$

◁

Por 4.10 vem que:

$$p_{1.1} = \frac{L'p_1 + L\pi - R_1}{L' + L}$$

e

$$p_{1.2} = p_1 - p_{1.1}$$
$$p_{2.1} = \pi - p_{1.1}$$

$$\text{então } R_1 = L'p_{1.2} + Lp_{2.1}$$

O risco 2 é dado por:

$$R_2 = L'p_{.12} + Lp_{.21}$$

Demonstração:

Por 4.9 vem que $R_2 = L'(1-\pi)(1-E_2) + L\pi(1-S_2)$, e por definição de sensibilidade e especificidade vem que:

S_2 =P(teste 2 positivo|g.s. positivo)= $\frac{P(teste2positivo,g.s.positivo)}{P(g.s.positivo)} = \frac{p_{.11}}{\pi}$

E_2 =P(teste 2 negativo|g.s. negativo)= $\frac{P(teste2negativo,g.s.negativo)}{P(g.s.negativo)} = \frac{p_{.22}}{1-\pi}$

substituindo em R_2:

$R_2 = L'(1-\pi)(1-\frac{p_{.22}}{1-\pi}) + L\pi(1-\frac{p_{.11}}{\pi}) =$

$= L'(1-\pi-p_{.22}) + L(\pi - p_{.11})$

e sabendo que $1-\pi-p_{.22} = p_{.12}$ e $\pi - p_{.11} = p_{.21}$, assim:

$R_2 = L'p_{.12} + Lp_{.21}$

◁

As estimativas para p_1, π, R_1 e R_2 são dadas por:

$$\widehat{p_1} = \frac{l_{1..}}{n}$$
$$\widehat{\pi} = \frac{n_1}{n}$$
e
$$\widehat{R}_1 = L'\frac{l_{1.2}}{n} + L\frac{l_{2.1}}{n}$$

$$\widehat{R}_2 = L'\frac{l_{.12}}{n} + L\frac{l_{.12}}{n}$$

Para calcular os riscos é suficiente conhecer a razão $\frac{L'}{L}$.

A variância do risco 1 é dada por:

$$var(\widehat{R}_1) = ((L')^2 p_{1.2} + L^2 p_{2.1} - R_1^2)/n$$

De um modo semelhante calculam-se as estimativas de e $var(\widehat{R}_2)$ para o teste de diagnóstico 2.

A covariância dos dois riscos é dada pela expressão:

$$cov(\widehat{R}_1, \widehat{R}_2) = ((L')^2 p_{112} + L^2 p_{221} - R_1 R_2)/n$$

As estimativas de $var(\widehat{R}_1)$ e $var(\widehat{R}_2)$ são dadas por:

$$\widehat{var}(\widehat{R}_1) = ((L')^2 \frac{l_{1.2}}{n} + L^2 \frac{l_{2.1}}{n} - \widehat{R}_1^2)/n$$

$$\widehat{var}(\widehat{R}_2) = ((L')^2 \frac{l_{.12}}{n} + L^2 \frac{l_{.21}}{n} - \widehat{R}_2^2)/n$$

logo, a estimativa da covariância é dada por:

$$\widehat{cov}(\widehat{R}_1, \widehat{R}_2) = ((L')^2 \frac{l_{112}}{n} + L^2 \frac{l_{221}}{n} - \widehat{R}_1 \widehat{R}_2)/n$$

Para testar a hipótese de igualdade de riscos nos dois testes de diagnóstico Bloch [4] define a seguinte estatística de teste:

$$Z = \frac{R_1 - R_2}{\sqrt{var\widehat{R}_1 + var(\widehat{R}_2) - 2cov(\widehat{R}_1, \widehat{R}_2)}} \tag{4.11}$$

Esta estatística de teste sob a hipótes nula, segue aproximadamente uma distribuição Gaussiana padrão.

Exemplo de Aplicação

Vamos considerar as amostras (fictícias) globais para dois testes de diagnóstico (tabela 4.3). Vamos supor que estamos interessados no caso $L' = L$, isto é, os custos associados a falsos positivos e falsos negativos são os mesmos. Para testar a hipótese de igualdade de riscos, vamos utilizar a estatística 4.11, em primeiro lugar temos de calcular as estimativas de R_1 e R_2.

	Teste 1 Positivo (1)	**Teste 1 Negativo (2)**	
Teste 2 Positivo (1)	$l_{111} = 169$	$l_{211} = 7$	180
	$l_{112} = 4$	$l_{212} = 0$	
Teste 2 Negativo (2)	$l_{121} = 8$	$l_{221} = 40$	222
	$l_{122} = 12$	$l_{222} = 162$	
	193	209	402

Tabela 4.3: Amostra fictícia global de dois testes de diagnóstico

Como $L' = L$, vem que as estimativas dos riscos são dados por:

$$\widehat{R}_1 = L(\frac{l_{1.2} + l_{2.1}}{n}) = L(\frac{16 + 47}{402}) = 0.1567L$$

$$\widehat{R}_2 = L(\frac{l_{.12} + l_{.21}}{n}) = L(\frac{4 + 48}{402}) = 0.1294L$$

E as estimativas das variâncias por:

$$\widehat{var}(\widehat{R}_1) = 0.1321L^2$$

$$\widehat{var}(\widehat{R}_2) = 0.11266L^2$$

$$\widehat{cov}(\widehat{R}_1, \widehat{R}_2) = 0.08922L^2$$

Então o valor da estatística de teste é $z = 2.089$, e para um valor-p=0.017, o risco usando o teste 2 é significativamente inferior ao de usar o teste 1.

4.3 Área Parcial Abaixo da Curva ROC (APAC)

A AUC é muito utilizada como índice de avaliação da performance de um teste de diagnóstico, no entanto não é um bom índice de avaliação quando uma elevada sensibilidade é exigida clinicamente (exemplo de um teste de diagnóstico com elevada sensibilidade é a mamografia, que ajuda a reduzir a mortalidade do cancro de peito identificando muito cedo os cancros que podem ser de imediato tratados). Alguns pressupostos clínicos exigem que alguns testes de diagnóstico tenham uma elevada sensibilidade. Para estes testes só alguns pontos de corte da curva são clinicamente aceitáveis, os outros pontos de corte com baixa sensibilidade são clinicamente irrevelantes. A área parcial abaixo da curva (APAC) ROC avalia a performance de um teste de diagnóstico apenas para uma região de interesse da sensibilidade [41, 27]. Pode dar-se o caso de ao compararmos duas curvas ROC elas não revelarem diferenças significativas quanto à AUC e no entanto revelarem diferenças para uma dada APAC.

Quando se comparam duas curvas e elas se cruzam, a AUC também não é um bom índice de avaliação. Se duas curvas ROC se cruzam num ponto onde a sensibilidade é inferior à sensibilidade exigida pelo clínico (S_0), então a APAC pode ser utilizada para comparar as perfomances dos testes. Se as curvas se cruzam num ponto onde a a sensibilidade é superior a S_0, a APAC tem o mesmo problema que a AUC. Mesmo que as curvas não se cruzem pode haver diferenças significativas para um determinada região que o índice global, AUC, não detecta.

Vamos abordar os trabalhos de Nishikawa *et al* [21], que propõem uma APAC derivada do modelo binormal convencional. Esta APAC paramétrica, sumaria a performance de um teste de diagnóstico numa região da curva ROC com elevada sensibilidade.

Zhang *et al* [46] propõem um algoritmo não-paramétrico para comparar a APAC para amostras emparelhadas. Para a estimação da variância e covariância da APAC baseiam-se nos algoritmos de DeLong *et al* [11] e Hanley *et al* [18]. Desenvolveram um programa em "SAS" para implementar o algoritmo (disponível pelo autor), mas também, para amostras pequenas, pode-se

desenvolver um programa numa folha de cálculo.

Empiricamente a análise da APAC por métodos paramétricos e não-paramétricos pode diferir muito mais do que a análise da AUC por estes dois métodos e é especialmente notório quando os testes de diagnóstico são qualitativos [46].

4.3.1 Cálculo da APAC

Vamos assumir que os dados provêm de um modelo binormal (ver secção 2.3.3), a sensibilidade e a proporção de falsos positivos são especificadoas a partir dos parâmetros da curva ROC binormal, a e b:

$$S(c) = \Phi(a - bc)$$
$$1 - E(c) = \Phi(-c)$$

onde c respresenta o ponto de corte que define a positividade da doença e

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2} dt$$

Cálculo da APAC

Nishikawa *et al* [21] definem a APAC por $A'_{z(S_0)}$, que representa a área abaixo da curva ROC, mas acima de um ponto de corte para uma sensibilidade pré-definida S_0. S_0 representa a fracção de verdadeiros positivos para um ponto de corte definido num contexto clínico. Este parâmetro livre, representa o mínimo aceitável para a sensibilidade num teste de diagnóstico numa determinada situação.

A APAC é defida pela expressão:

$$A'_{z(S_0)} = \frac{\int_{S_0}^{1}(1-(1-E)S)dS}{1-S_0} \tag{4.12}$$

A APAC tem propriedades semelhantes à AUC e pode ser interpretada como a média da epecificidade para todos os valores da sensibilidade entre S_0 e 1. O intervalo de variação da APAC é definido em:

$$\frac{1}{2}(1-S_0) \leq A'_{Z(S_0)} \leq 1$$

Contrariamente à AUC, a APAC pode ter valores inferiores a 0.5, e quando $S_0 = 0$ vem que APAC=AUC.

Aplicando uma transformação de variável a (4.12) e considerando o modelo binormal a APAC pode ser definida por:

$$A'_{z(S_0)} = 1 - \frac{b}{\sqrt{2\pi}(1-S_0)} \int_{-\infty}^{S_0} \Phi(x) e^{-\frac{(bx-a)^2}{2}} dx \tag{4.13}$$

onde S_0 é o parâmetro fixo e a e b são calculados computacionalmente, por exemplo pelo "ROCKIT", e a APAC pode ser calculada numericamente.

4.3.2 Comparação das APAC de duas curvas ROC

Como a APAC é conceptualmente semelhante à AUC, os testes para a diferença das APACs são semelhantes aos testes para comparar as diferenças das

AUCs.

Para testar $H_0 : \theta'_{z1(S_0)} = \theta'_{z2(S_0)}$, a estatística de teste é dada por [21]:

$$Z = \frac{A'_{z1(S_0)} - A'_{z2(S_0)}}{\sqrt{var(A'_{z1(s_0)} - A'_{z2(S_0)})}} \tag{4.14}$$

onde $var(A'_{z1(S_0)} - A'_{z2(S_0)}) = var(A'_{z1(S_0)}) + var(A'_{z2(S_0)}) - 2cov(A'_{z1(S_0)}, A'_{z2(S_0)})$

A variância da APAC, é uma função de S_0 e dos parâmetros binormais a e b:

$$var(A'_{z(S_0)}) = (\frac{\partial A'_{z(S_0)}}{\partial a})^2 var(a) + (\frac{\partial A'_{z(S_0)}}{\partial b})^2 var(b) + 2(\frac{\partial A'_{z(S_0)}}{\partial a})(\frac{\partial A'_{z(S_0)}}{\partial b}) cov(a, b)$$

onde:

$$\frac{\partial A'_{z(S_0)}}{\partial a} = \frac{e^{-\frac{a^2}{2(1+b^2)}}}{(1 - S_0)\sqrt{2\pi(1 + b^2)}} \times (1 - \Phi(\lambda))$$

$$\frac{\partial A'_{z(S_0)}}{\partial b} = \frac{e^{-\frac{a^2}{2(1+b^2)} - \frac{\lambda^2}{2}}}{2\pi(1 + b^2)(1 - S_0)} - \frac{abe^{-\frac{a^2}{2(1+b^2)}}}{\sqrt{2\pi}(1 + b^2)^{\frac{3}{2}}(1 - S_0)} \times (1 - \Phi(\lambda))$$

com

$\Phi(\lambda) = \frac{\sqrt{1+b^2}}{b}\Phi^{-1}(S_0) - \frac{a}{b\sqrt{1+b^2}}$

Se as duas curvas são estimadas a partir de duas amostras independentes, então $cov(A'_{z1(S_0)}, A'_{z2(S_0)}) = 0$, mas se as amostras são emparelhadas a covariância é dada por:

$$cov(A'_{z1(S_0)}, A'_{z2(S_0)}) = (\frac{\partial A'_{z1(S_0)}}{a_1})(\frac{\partial A'_{z2(S_0)}}{a_2})cov(a_1, a_2)+$$
$$+(\frac{\partial A'_{z1(S_0)}}{a_1})(\frac{\partial A'_{z2(S_0)}}{b_2})cov(a_1, b_2) + (\frac{\partial A'_{z1(S_0)}}{b_1})(\frac{\partial A'_{z2(S_0)}}{a_2})cov(b_1, a_2)+$$
$$+(\frac{\partial A'_{z1(S_0)}}{b_1})(\frac{\partial A'_{z2(S_0)}}{b_2})cov(b_1, b_2)$$

Podemos calcular as covariâncias dos parâmetros binormais através dos programas "CORROC" para dados de natureza discreta e "CLABROC" para dados de natureza contínua.

A estatística de teste para amostras grandes, segue uma distribuição assintóticamente Normal padrão. No entanto, a maior parte das vezes as amostras disponíveis são geralmente pequenas. Nishikawa *et al* [21] propuseram a seguinte transformação não linear para a APAC:

$$\widehat{\vartheta} = \frac{1}{2}ln(\frac{1 + A'_{z(S_0)}}{1 - A'_{z(S_0)}})$$

Então a estatística de teste (4.14) passa a ser:

$$Z = \frac{\widehat{\vartheta}_1 - \widehat{\vartheta}_2}{\sqrt{var(\widehat{\vartheta}_1) + var(\widehat{\vartheta}_2) - 2cov(\widehat{\vartheta}_1, \widehat{\vartheta}_2)}}$$

onde $var(\widehat{\vartheta}) = \frac{1}{(1-(A'_{z(S_0)})^2)}var(A'_{z(S_0)})$

e $cov(\widehat{\vartheta}_1, \widehat{\vartheta}_2) = \frac{1}{1-(A'_{z1(S_0)})^2}(1 - (A'_{z2(S_0)})^2)cov(A'_{z1(S_0)}, A'_{z2(S_0)})$

Sob a hipótese nula, esta estatística de teste segue aproximadamente uma distribuição Normal padrão.

Capítulo 5

Ponto de Corte

A escolha do ponto de corte óptimo é também um dos principais objectivos quando se estuda a eficácia de um teste de diagnóstico. Com efeito, este ponto de corte óptimo será usado no futuro para decidir se o resultado do teste de diagnóstico é positivo ou negativo. Se o teste de diagnóstico for positivo e posteriormente se verificar que o indivíduo não está doente então diz-se que houve um "falso alarme", se o teste de diagnóstico for positivo e se verificar que o indivíduo tem de facto a doença diz-se então que houve um "sucesso". Pode-se limitar os falsos alarmes, mas em detrimento dos sucessos. A questão da escolha do ponto de corte para diagnosticar a doença não tem uma resposta simples, já que existem muitos critérios que servem de base para a decisão, tais como:

- custos financeiros que estão directa ou indirectamente ligados ao tratamento da doença (presente ou não) e ao não tratamento da doença
- custos associados a futuras investigações
- desconforto dos pacientes causado pelos tratamentos
- mortalidade associada ao tratamento ou não tratamento.

Parece ser lógico que quando o custo associado à falha de um diagnóstico é elevado e o tratamento (mesmo um tratamento inapropriado a um indivíduo não doente) é seguro, então deve-se mover o ponto de corte para a esquerda (ver figura do movimento do ponto de corte para a esquerda na secção 1) da curva ROC, onde temos uma elevada proporção de verdadeiros positivos (maioria dos verdadeiros positivos são tratados, assim como muitos falsos positivos). Se os riscos de terapia são elevados e esta não proporciona muita ajuda, devemos colocar o ponto de corte para a direita (ver figura do movimento do ponto de corte para a direita na secção 1) da curva ROC, onde se vai obter uma baixa proporção de verdadeiros positivos, mas por outro lado também não se irá aplicar um tratamento pouco seguro a indivíduos não doentes.

Neste trabalho vamos abordar duas situações para o cálculo do ponto de corte. A primeira abordagen é não-paramétrica e vai ter em consideração os custos associados à realização de um teste de diagnóstico assim como os custos associados às decisões resultantes (exemplo falsos positivos, verdadeiros positivos, etc). A segunda abordagem tem uma vertente paramétrica e uma não-paramétrica, e é adequada quando estamos na presença de um teste de diagnóstico quantitativo.

5.1 Cálculo do ponto de corte considerando custos associados ao teste de diagnóstico

Quando se realiza um teste de diagnóstico, além de um custo inicial (C_0) há sempre custos adicionais associados aos vários resultados possíveis desse teste. Assim a função custo de um teste de diagnóstico pode ser representada através da tabela 5.1.

	Doença Presente	**Doença Ausente**
Teste Positivo	C_{VP}	C_{FP}
Teste Negativo	C_{FN}	C_{VN}

Tabela 5.1: Tabela da função custo

Onde:

C_{VP} = custo associado aos verdadeiros positivos

C_{VN} = custo associado aos verdadeiros negativos

C_{FP} = custo associado aos falsos positivos

C_{FN} = custo associado aos falsos negativos

C_0 = custo associado à realização do teste

Deste modo, o custo associado a um teste de diagnóstico é uma variável aleatória cujo valor esperado, $E(Custo)$, é facilmente calculado através do conhecimento das probabilidades associadas aos diferentes resultados do

teste. Usando a notação introduzida na tabela 1.4 no capítulo 1, esse valor pode ser expresso por [22]:

$$E(Custo) = C_0 + C_{VP} \times P(VP) + C_{VN} \times P(VN) + C_{FP} \times P(FP) + C_{FN} \times P(FN) \tag{5.1}$$

Considerando as tabelas 1.1 e 1.2 no capítulo 1 vem que:

$P(VP) = P(D) \times P(+ \mid D) = P(D) \times PVP$

$P(VN) = P(\bar{D}) \times P(- \mid \bar{D}) = P(\bar{D}) \times PVN$

$P(FP) = P(\bar{D}) \times P(+ \mid \bar{D}) = P(\bar{D}) \times PFP$

$P(FN) = P(D) \times P(- \mid D) = P(D) \times PFN$

Substituindo em (5.1):

$$E(Custo) = C_0 + C_{VP} \times P(D) \times PVP + C_{VN} \times P(\bar{D}) \times PVN +$$

$$+ C_{FP} \times P(\bar{D}) \times PFP + C_{FN} \times P(D) \times PFN \tag{5.2}$$

Sabendo que $PVN = 1 - PFP$ e $PFN = 1 - PVP$ e substituindo em (5.2) vem:

$$E(Custo) = C_0 + C_{(VP)} \times P(D) \times PVP + C_{VN} \times P(\bar{D}) \times (1 - PFP) +$$

$$+ C_{FP} \times P(\bar{D}) \times PFP + C_{FN} \times P(D) \times (1 - PVP) \tag{5.3}$$

que é equivalente a ter:

$$E(Custo) = C_0 + PVP \times P(D) \times (C_{VP} - C_{FN}) + \\ + PFP \times P(\bar{D}) \times (C_{FP} - C_{VN}) + C_{VN} \times P(\bar{D}) + C_{FN} \times P(D) \quad (5.4)$$

Ao analisarmos a equação (5.4) pode-se observar que a dependência do $E(Custo)$ pela PVP e pela PFP, isto é pelas coordenadas da curva ROC, implica que o custo médio depende do ponto de corte definido. Fazendo variar este ponto de corte varia o custo médio. O custo óptimo será alcançado quando o $E(Custo)$ é mínimo. Uma vez que se consegue expressar PVP (sensibilidade) em função de PFP (1-especificidade) através da curva ROC, $\varphi(PFP) = PVP$, leva à seguinte equação:

$$E(Custo) = C_0 + \varphi(PFP) \times P(D) \times (C_{VP} - C_{FN}) +$$

$$+ PFP \times P(\bar{D}) \times (C_{FP} - C_{VN}) + C_{VN} \times P(\bar{D}) + C_{FN} \times P(D) \quad (5.5)$$

Derivando (5.5) em ordem a PFP tem-se:

$$\frac{\partial E(Custo)}{\partial PFP} = \frac{\partial \varphi}{\partial PFP} \times P(D) \times (C_{VP} - C_{FN}) - P(\bar{D}) \times (C_{VN} - C_{FP}) \quad (5.6)$$

Igualando (5.6) a zero, obtém-se:

$$\frac{\partial \varphi}{\partial PFP} \times P(D) \times (C_{VP} - C_{FN}) = -P(\bar{D}) \times (C_{FP} - C_{VN}) \quad (5.7)$$

Ou seja

$$\frac{\partial \varphi}{\partial PFP} = \frac{P(\bar{D}) \times (C_{FP} - C_{VN})}{P(D) \times (C_{FN} - C_{VP})} = m \tag{5.8}$$

A equação (5.8) é a equação diferencial que define o declive da recta tangente à curva ROC no ponto cujo custo é óptimo.

Se o teste de diagnóstico for qualitativo, poder-se-á calcular o valor médio do custo a partir da expressão (5.5) para cada ponto de corte, e escolher o que tiver o valor médio do custo mínimo. Se o teste de diagnóstico for quantitativo, é necessário ter o conhecimento da função da curva ROC, φ, quando a curva é estimada a partir de um método não paramétrico, não temos o conhecimento da função da curva ROC, então de uma forma equivalente à situação de termos um teste qualitativo, vamos calcular o valor do custo médio a partir da expressão (5.5) para todos os pontos que serviram para a construção da curva, e escolher o que tem custo mínimo.

Ao calcular o ponto de corte deste modo há que ter em consideração o seguinte:

- Quando a doença é rara, $\frac{P(\bar{D})}{P(D)}$ será elevado, então deve-se mudar o ponto de corte para a parte mais próxima do vértice (0,0) do plano unitário, onde o declive (5.8) será grande. Para doenças muito raras, ter falsos positivos é muito grave, deve-se minimizar os falsos positivos, mesmo à custa de não se diagnosticar verdadeiros positivos.

- Quando a doença em causa é comum, deve-se mudar o ponto de corte para mais próximo do vértice (1,1) do plano unitário. Nesta situação a maior parte dos negativos são falsos negativos.

- O declive da tangente à curva será tanto maior quanto maior for a diferença dos custos $C_{FP} - C_{VN}$ relativamente a $C_{FN} - C_{VP}$. Considere-se a seguinte situação prática, assume-se que para uma doença particular (tumor cerebral) obteve-se um teste positivo. Então tem-se que operar o indivíduo e encontrar o presumivel tumor. Se o teste for negativo

não se faz nada. Vai-se também assumir que a operação não é muito eficaz no tratamento (a probabilidade de morrer durante a operação é considerável). O custo de um falso positivo (operar o cérebro de um indivíduo sem tumor) é de facto superior relativamente ao custo de um verdadeiro negativo (não fazer nada). O declive é elevado, então deve-se mudar o ponto de corte para a esquerda da curva ROC.

- O caso oposto ao anterior será quando as consequências de um falso positivo são mínimas. Neste caso tem-se um grande benefício em tratar os doentes. Nesta situação deve-se mudar o ponto de corte para a direita da curva ROC.

Exemplo de Aplicação

Vamos determinar o ponto de corte óptimo para o modelo binormal. Neste caso temos o conhecimento da função da curva ROC.

A função da curva ROC no modelo binormal é dada pela expressão:

$\Phi^{-1}(S) = a + b\Phi^{-1}(1 - E)$

então o declive da recta tangente à curva é por definição $m = (\Phi^{-1}(S))'$, e pela definição da derivada da função inversa temos que:

$$m = \frac{1}{\Phi'(\frac{\mu_y - c}{\sigma_y})} = \frac{1}{\phi(\frac{\mu_y - c}{\sigma_y})} = \frac{1}{\frac{1}{\sqrt{2\pi}} e^{-1/2(\frac{\mu_y - c}{\sigma_y})^2}}$$

logo o ponto de corte é dado por:

$c = \mu_y - \sigma_y \sqrt{2ln(\frac{m}{\sqrt{2\pi}})}$

5.2 Cálculo do ponto de corte para testes de diagnóstico quantitativos

Helmut Schafer [36] propõe dois procedimentos para o cálculo do ponto de corte para um teste de diagnóstico quantitativo, numa situação em que é especificado à partida a especificidade e/ou sensibilidade para a escolha do ponto de corte. Dado que os valores do teste de diagnóstico obtidos a partir da amostras de doentes e não doentes não contêm informação relevante para a determinação da sensibilidade e da especificidade, não existe nenhuma razão formal para não se especificar estes dois valores antes de se analisarem os dados [36].

No procedimento A, é estipulado à partida a especificidade desejada; e no procedimento B, a escolha do ponto de corte depende da escolha da sensibilidade e da especificidade em simultâneo.

Como anteriormente X variável aleatória que representa os valores da variável resposta para a população de não doentes e Y a variável aleatória que representa os valores da variável resposta para a população de doentes, e consideram-se as seguintes notações:

- $z_\gamma = \phi^{-1}(\gamma), 0 < \gamma < 1$, representa o quantil de probabilidade γ da Gaussiana padrão
- f_X é a função de densidade de probabilidade da variável X; se esta tiver distribuição Gaussiana, os parâmetros são μ_x e σ_x
- g_Y é a f.d.p. da variável Y; se tiver distribuição Gaussiana, os parâmetros são μ_y e σ_y
- $E(c)$ e $S(c)$ são os valores da especificidade e da sensibilidade para o ponto de corte c, e E e S são os seus valores pré-fixados

- χ_E e ξ_S são respectivamente os quantis de probabilidade E e S relativamente às correspondentes distribuições de X e Y, são definidos respectivamente por $E(\chi_E) = E$ e $S(\xi_S) = S$

Para duas amostras independentes de X e Y seja ainda:

- n dimensão da amostra de não doentes e m dimensão da amostra de doentes
- Para um dado valor c, $\widehat{E}(c)$ e $\widehat{S}(c)$ são os estimadores da especificidade e da sensibilidade
- As médias e desvios-padrão empíricos são dados por $\bar{x}, \bar{y}, s_x, s_y$
- $\widehat{f}_X$ e $\widehat{g}_Y$ são os estimadores consistentes das respectivas funções densidade
- os estimadores paramétricos para os quantis χ_E e ξ_S são dados respectivamente por: $X_E = \bar{X} + s_X z_E$ e $Y_S = \bar{Y} - s_Y z_S$; e os estimadores não-paramétricos são dados respectivamente por: $X_E = nE$ (ordem na amostra X) e $Y_S = m(1 - S)$ (ordem na amostra Y)
- As expressões para as variâncias no caso Gaussiano são dadas por $\sigma^2_{\chi_E} = \frac{\sigma^2_x}{n}(1 + \frac{z^2_E}{2})$ e $\sigma^2_{\xi_S} = \frac{\sigma^2_y}{m}(1 + \frac{z^2_S}{2})$, e no caso não-paramétrico são dadas por $\sigma^2_{\chi_E} = \frac{E(1-E)}{f^2_X(\chi_E)n}$ e $\sigma^2_{\xi_S} = \frac{S(1-S)}{f^2_Y(\xi_S)m}$
- As estimativas de $S^2_{\chi_E}$ e $S^2_{\xi_S}$ são obtidos através de s_x e s_y ou $\widehat{f}_X(X_E)$ e $\widehat{f}_Y(Y_S)$

Procedimento A para o cálculo do ponto de corte

1. Seleccionar a especificidade E desejada
2. O ponto de corte c é definido a partir do limite superior do intervalo de confiança para χ_E, com uma confiança de $\sqrt{1-\alpha}$. Uma solução aproximada é dada por: $c = \chi_E + z_{\sqrt{1-\alpha}} s_{\chi_E}$

3. O limite inferior de confiança $\sqrt{1-\alpha}$, $\widehat{S}_l(c)$, para a sensibilidade $S(c)$, é calculado a partir da amostra Y. No caso não paramétrico o método usado por Clopper e Pearson [8] pode ser usado para uma solução exacta, ou uma aproximação Gaussiana dada por $\widehat{S}_l(c) = \widehat{S}(c) - z_{\sqrt{1-\alpha}}(\widehat{S}(c)\frac{1-\widehat{S}(c)}{m})^{1/2}$

A escolha de $\sqrt{1-\alpha}$ assegura uma confiança global de probabilidae $1-\alpha$ para a condição "$E(c) \geq E$ e $S(c) \geq \widehat{S}_l(c)$". O limite de confiança $\widehat{S}_l(c)$ determina quando o resultado da regra de classificação é satisfatória ou não. Formalmente, $\widehat{S}_l(c)$ é comparado com o valor mínimo de S considerado satisfatório para a aplicação da E. No entanto para aplicar o procedimneto A, não é necessário dar a sensibilidade S em avanço. Quando isto acontece, nenhum limite de confiança $\widehat{S}_l(c)$ é necessário, mas apenas a decisão "$E(c) \geq E$ e $S(c) \geq S$". Se a condição não for satisfeita, o procedimento B é mais eficiente.

Procedimento B para o cálculo do ponto de corte c

1. Definir os valores mínimos requeridos para a sensibilidade (S) e para a especificidade (E)
2. O ponto de corte é dado pela expressão $c = w\chi_E + (1-w)\xi_S$, onde os pesos são dados por $w = \frac{S_{\xi_S}}{S_{\chi_E} + S_{\xi_S}}$
3. A estatística de teste de Greenhouse e Mantel:

$$T = \frac{Y_S - X_E}{\sqrt{S^2_{X_S} + S^2_{Y_E}}}$$

é usada para decidir quando ($T > t$) ou não ($T \leq t$)) o ponto de corte seleccionado conduz à condição $E(c) \geq E$ e $S(c) \geq S$.

O problema agora é determinar qual o valor crítico de t tal que o erro de tipo I é inferior a α. Schafer [36], propõe uma aproximação assintótica

para o valor crítico:

$$t = -2z_{\sqrt{\frac{\alpha}{2}}}$$

Não é correcto aplicar o teste de Greenhouse e Mantel com um limite crítico de $t = z_{1-\alpha}$, uma vez que este iria aumentar o erro de tipo I [36].

Exemplo de Aplicação

Vamos considerar o teste de diagnóstico IgE Total descrito no capítulo 1.2. Uma vez que o logaritmo dos valores do teste de diagnóstico para as duas amostras seguem uma distribuição Gaussiana, vamos utilizar a abordagem paramétrica.

Procedimento A:

1. Vamos considerar uma especificidade $E = 0.95$.

2. Para calcular o ponto de corte c vamos fazer os seguintes cálculos intermédios:

$X_{0.95} = 3.85 + 1.41 + 1.645 = 6.16945$

$S^2_{X_{0.95}} = (\frac{1.412^2}{60}(1 + \frac{1.6452^2}{2}) = 0.07797$

Então o ponto de corte c é:

$c = 6.16945 + 2.576 * 0.27923 = 6.887$

Procedimento B:

1. Vamos considerar uma sensibilidade de $S = 0.98$ e uma especificidade de $E = 0.95$

2. Para calcular c vamos fazer os seguintes cálculos intermédios:

$Y_{0.98} = 5.33 - 1.163 * 2.26 = 2.70162$

$S^2_{Y_{0.98}} = \frac{1.163^2}{51} * (1 + \frac{2.26^2}{2}) = 0.09425$

$w = \frac{0.307}{0.27923+0.307} = 0.5237$

Então o ponto de corte é:

$c = 0.5237 * 6.16945 + 0.4763 * 2.70162 = 4.518$

3. Vamos agora calcular o valor da estatística de teste de Greenhouse e Mantel:

$t = \frac{2.70162-6.16945}{\sqrt{0.09425+0.07797}} = -8.356$

$t = -2z_{\frac{0.05}{2}} = 2$

Então o ponto de corte seleccionado não satisfaz a condição $S(c) \geq S$ e $E(c) \geq E$.

Capítulo 6

Conclusões e Considerações Finais

O objectivo principal deste trabalho foi caracterizar as curvas ROC como instrumento na análise estatística de testes de diagnóstico.

As curvas ROC são um instrumento de análise da performance de testes de diagnóstico que discriminem entre duas populações. A construção destas curvas baseiam-se em duas quantidades, a sensibilidade e a especificidade. Não dependem de um ponto de corte específico, uma vez que ela é construida com base em todos os pontos de corte possíveis. É invariante a transformações de escala da variável resposta e é de fácil interpretação geométrica.

De entre os métodos de estimação da curva ROC, os paramétricos são os mais aplicados e explorados particularmente o modelo binormal. Dos métodos não paramétricos o método empírico é o mais simples de aplicar, mas por outro lado produz curvas irregulares que resultam numa sub-estimação da área abaixo da curva. O método *kernel* revelou-se o menos optimista, isto é, foi o que apresentou um valor da área abaixo da curva mais baixo. Quanto à escolha da amplitude para a construção da função *kernel*, a proposta de Shapiro *et al* [48] parece ser mais adequada quando se utiliza a função *kernel* Gaussiana, no entanto não se revelou óptima para esta última. Quando são determinadas as amplitudes óptimas para as funções *kernel* Gaussiana e biweight, as respectivas curvas ROC não revelam diferenças significativas.

A curva ROC estimada pelo modelo binormal e representada no plano binormal, é uma recta. O modelo binormal próprio é uma alternativa ao modelo convencional, quando neste último as curvas ROC resultantes são degeneradas.

A área abaixo da curva (AUC) é o índice mais utilizado para avaliar a performance global de um teste de diagnóstico. Este índice varia de 0.5 a 1, e quanto mais próximo de 1, melhor é o poder discriminativo do teste. Quanto à comparação de testes de diagnóstico a AUC não é a mais adequada quando estamos na presença de curvas que se cruzam, ou quando alguns pontos de corte não são clinicamente aceitáveis, então a alternativa é a comparação da área parcial abaixo da curva ROC (APAC). Bloch [4] propõe ainda a comparação de riscos associados aos testes. Quando comparamos as AUC de testes de diagnósticos, estas podem não revelar diferenças significativas e no entanto existirem diferenças de riscos.

A escolha do ponto de corte óptimo num teste de diagnóstico é uma opção muito importante a tomar, pois os resultados do diagnóstico de futuros indivíduos dependem desta decisão. Neste trabalho abordámos duas situações, uma envolvendo custos associados à realização do teste de diagnóstico e outra situação em que é especificado à partida a e especificidade e/ou sensibilidade à partida. O primeiro é um processo moroso quando a variável resposta é contínua e se não temos a função da curva ROC, pois temos de calcular o custo médio para todos os pontos de corte possíveis e seleccionar aquele que tiver o menor custo médio.

O tema relativo às curvas ROC, sugere a aplicação de várias técnicas estatísticas, é efectivamente uma área que tem muito por explorar, e deixamos algumas técnicas que não foram abordadas neste trabalho.

A interpretação de um teste de diagnóstico pode ser feita por vários estudos independentes. A curva SROC (*summary receiver operating characteristic*) [33, 43] tem por objectivo representar a relação da PVP e PFP ao longo dos estudos, mesmo que estes utilizem diferentes pontos de corte.

Recentemente o interesse vai para além da determinação da exactidão de um teste de diagnóstico. Os investigadores estão interessados em determinar

quais são os factores que afectam a exactidão de um teste de diagnóstico. Ao se determinarem estes factores é possível identificar populações onde o teste é mais ou menos exacto. Isto é possível através da análise de regressão [2, 25, 24]. Considera-se um modelo tal que a curva ROC é uma função de covariáveis e estas podem ser comuns a todos os indivíduos ou só especificamente para indivíduos doentes.

Investigadores usam as curvas ROC para descrever a performance dos sistemas de diagnóstico que podem ser qualitativos ou contínuos. A análise ROC tradicionalmente assume que um teste de diagnóstico envolve a distinção entre duas condições mutuamente exclusivas (doença presente ou ausente). Mas os clínicos normalmente enfrentam situações em que têm de decidir entre três ou mais alternativas de diagnóstico. Por exemplo, nas leituras das mamografias, o radiologista decide se existem nódulos benignos, nódulos malignos ou se não existem nódulos; o psiquiatra que avalia um indivíduo psicótico, muitas vezes tem de decidir se o indivíduo sofre de uma psicose afectiva, esquizofrenia ou uma desordem ilusional. Muitos destes problemas de diagnóstico são resumidos em termos dicotómicos (doença presente ou ausente), mas esta caracterização corre o risco de distorcer problemas complexos e difíceis.

Surge então uma nova metodologia para avaliar a exactidão de testes de diagnóstico que classificam indivíduos em três categorias, a chamada Superfície ROC [45, 34].

Yang *et al* [45] propõem a generalização da curva ROC para analisar o caso em que um teste de diagnóstico classifica em três categorias: doente, não doente e indeterminado. Este método consiste em representar a PVP *versus* PFP e PI (proporção de indeterminados).

Sejam X, Y e S as variáveis aleatórias que representam os valores da variável resposta para as populações de não doentes, doentes e indeterminados respectivamente. Para cada par de pontos de corte (c, s), definem-se as PVP, PFP e PI da seguinte maneira:

$$PVP = P(Y > c, S > s) = P(Y > c|S > s)P(S > s)$$

$$PFP = P(X > c, S > s) = P(X > c|S > s)P(S > s)$$

$PI = P(S > s)$

onde $0 \leq PFP \leq P(S > s) = 1 - PI$

Cada selecção do par (c, s) conduz a um valor de PVP, PFP e PI, que quando representado num espaço tridimensional formam um ponto. Todos os pares de pontos de corte possíveis, formam uma superfície no espaço. Designamos esta superfície por superfície ROC de um teste de diagnóstico (figura 6.1).

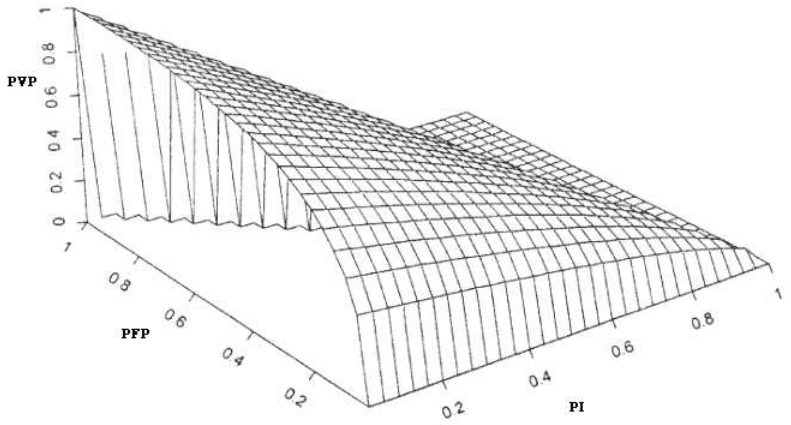

Figura 6.1: Superfície ROC

A escolha ideal dos pontos de corte (c, s), é aquela que maximiza a PVP e minimiza a PFP e PI. No entanto, como se pode ver no gráfico 6.1, a PVP aumenta à medida que a PFP aumenta, isto é, que a PVN diminui. No entanto o decréscimo da PI resulta no aumento da PVP, que por sua vez aumenta a PFP. Isto deve-se ao facto da PFP ter uma relação funcional semelhante com a PI e com a PVP.

O facto da superfície ROC ter uma elevação no canto (1,0,0) do gráfico 6.1 indica uma elevada sensibilidade e pequenas PFP e PI. Podemos identificar semelhanças com a curva ROC. Assim como a área abaixo da curva ROC

traduz um índice de avaliação da performance de um teste de diagnóstico, o volume abaixo da superfície ROC também traduz uma medida da performance global de um teste. É intuitivo que quanto maior for o volume abaixo da superfície melhor é a capacidade de discriminação do teste de diagnóstico.

Podemos estender as propriedades da curva ROC para a superfície ROC, e, em teoria, não há razão para que estes métodos não sejam estendidos a problemas de decisão que envolvam quatro ou mais possibilidades de diagnóstico. Dados multidimensionais serão difíceis de representar em gráficos bididimensionais, no entanto o cálculo da avaliação da performance de um teste de diagnóstico é um desafio!

O contributo da evolução tecnológica na investigação, é inquantificável. Assim, deixamos algumas sugestões de programas com aplicação das curvas ROC e respectivos *sites* para eventuais *donwloads*.

- Programas criados por Charles Metz:

  ROCKIT: Calcula as estimativas de máxima verosimilhança dos parâmetros do modelo binormal convencional, para dados de natureza ordinal ou contínua. Compara testes de diagnóstico para amostras independentes, emparelhadas e parcialmente emparelhadas. Este programa substitui o ROCFIT, LABROC1, INDROC, CORROC2 e CLABROC.

  PLOTROC.xls: Trata-se de uma macro em Excel para construir a curva ROC no plano unitário para o modelo binormal convencional.

  ROCFIT: Estima os parâmetros do modelo binormal convencional pelo método de máxima verosimilhança para dados em escala ordinal.

  LABROC1: Estima os parâmetros do modelo binormal convencional pelo método de máxima verosimilhança para dados contínuos.

  CORROC2: Compara dois testes de diagnóstico emparelhados para dados em escala ordinal

CLABROC: Compara dois testes de diagnóstico emparelhados para dados contínuos.

INDROC: Compara dois testes de diagnóstico independentes para dados em escala ordinal.

download a partir do site:
http://xray.bsd.uchicago.edu/krl/roc_soft.htm

- PROPROC é um programa desenvolvido por Pan e Metz para dados degenerados. Aplica uma abordagem paramétrica e pode ser aplicado a dados de natureza discreta ou contínua. *download* a partir do *site*:
ftp://minira.bsd.uchicago.edu/roc/ibmpc/

- MRMC, um programa de Kevin Berboum e inclui os programas RSCORE e BIGAMMA, para dados discretos degenerados, ambos aplicam uma abordagem paramétrica. *download* a partir do *site*:
ftp://perception.radiology.uiowa.edu/

- MULTIVARIATEROC.S, é um programa em S-PLUS desenvolvido por Hemant Ishwaran, e usa o método não paramétrico de DeLong para comparar testes de diagnóstico. O acesso pode ser feito a partir do *site*:
http://www.bio.ri.ccf.org/Resume/Pages/Ishwaran/multivariateRoc.s

- OBUMRM.FOR, é um programa em FORTRAN desenvolvido por Nancy Obuchowski, e implementa o método de Obuchowski-Rockette para analisar dados de múltiplos leitores de testes de diagnóstico. *download* a partir do *site*:
http://www.bio.ri.ccf.org/OBUMRM/OBUMRM.html

- Para uma abordagem Bayesiana da análise ROC, a partir do *site*:http://www-math.bgsu.edu/ albert/ord_book/chapter5/ pode-se ter acesso a um programa em MATLAB

- ROCNPA, um programa desenvolvido por Ana Cristina Braga, tem uma abordagem não paramétrica, compara mais de três testes de diagnóstico quer para amostras independentes quer para amostras emparelhadas.

"A intervenção da Estatística em Ciências Biomédicas tem um largo historial,
e a Estatística orgulha-se, a justo título, de ter contribuido
para a alteração do próprio paradigma
de o que é uma experiência científica."

in Introdução à Probabilidade e à Estatística, [35]

Parte I

Tabela de Coeficientes de Correlação entre Áreas

	0.700	0.725	0.750	0.775	0.800	0.825	0.850	0.875	0.900	0.925	0.950	0.975
0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.01	0.01	0.01	0.01	0.01
0.04	0.04	0.04	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.02	0.02	0.02
0.06	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.04	0.04	0.04	0.03	0.02
0.08	0.07	0.07	0.07	0.07	0.07	0.06	0.06	0.06	0.06	0.05	0.04	0.03
0.10	0.09	0.09	0.09	0.09	0.08	0.08	0.08	0.07	0.07	0.06	0.06	0.04
0.12	0.11	0.11	0.11	0.10	0.10	0.10	0.09	0.09	0.08	0.08	0.07	0.05
0.14	0.13	0.12	0.12	0.12	0.12	0.11	0.11	0.11	0.10	0.09	0.08	0.06
0.16	0.14	0.14	0.14	0.14	0.13	0.13	0.13	0.12	0.11	0.11	0.09	0.07
0.18	0.16	0.16	0.16	0.16	0.15	0.15	0.14	0.14	0.13	0.12	0.11	0.09
0.20	0.18	0.18	0.18	0.17	0.17	0.17	0.16	0.15	0.15	0.14	0.12	0.10
0.22	0.20	0.20	0.19	0.19	0.19	0.18	0.18	0.17	0.16	0.15	0.14	0.11
0.24	0.22	0.22	0.21	0.21	0.21	0.20	0.19	0.19	0.18	0.17	0.15	0.12
0.26	0.24	0.23	0.23	0.23	0.22	0.22	0.21	0.20	0.19	0.18	0.16	0.13
0.28	0.26	0.25	0.25	0.25	0.24	0.24	0.23	0.22	0.21	0.20	0.18	0.15
0.30	0.27	0.27	0.27	0.26	0.26	0.25	0.25	0.24	0.23	0.21	0.19	0.16
0.32	0.29	0.29	0.29	0.28	0.28	0.27	0.26	0.26	0.24	0.23	0.21	0.18
0.34	0.31	0.31	0.31	0.30	0.30	0.29	0.28	0.27	0.26	0.25	0.23	0.19
0.36	0.33	0.33	0.32	0.32	0.31	0.31	0.30	0.29	0.28	0.26	0.24	0.21
0.38	0.35	0.35	0.34	0.34	0.33	0.33	0.32	0.31	0.30	0.28	0.26	0.22
0.40	0.37	0.37	0.36	0.36	0.35	0.35	0.34	0.33	0.32	0.30	0.28	0.24
0.42	0.39	0.39	0.38	0.38	0.37	0.36	0.36	0.35	0.33	0.32	0.29	0.25
0.44	0.41	0.40	0.40	0.40	0.39	0.38	0.38	0.37	0.35	0.34	0.31	0.27
0.46	0.43	0.42	0.42	0.42	0.41	0.40	0.39	0.38	0.37	0.35	0.33	0.29
0.48	0.45	0.44	0.44	0.43	0.43	0.42	0.41	0.40	0.39	0.37	0.35	0.30
0.50	0.47	0.46	0.46	0.45	0.45	0.44	0.43	0.42	0.41	0.39	0.37	0.32
0.52	0.49	0.48	0.48	0.47	0.47	0.46	0.45	0.44	0.43	0.41	0.39	0.34
0.54	0.51	0.50	0.50	0.49	0.49	0.48	0.47	0.46	0.45	0.43	0.41	0.36
0.56	0.53	0.52	0.52	0.51	0.51	0.50	0.49	0.48	0.47	0.45	0.43	0.38
0.58	0.55	0.54	0.54	0.53	0.53	0.52	0.51	0.50	0.49	0.47	0.45	0.40
0.60	0.57	0.56	0.56	0.55	0.55	0.54	0.53	0.52	0.51	0.49	0.47	0.42
0.62	0.59	0.58	0.58	0.57	0.57	0.56	0.55	0.54	0.53	0.51	0.49	0.35
0.64	0.61	0.60	0.60	0.59	0.59	0.58	0.58	0.57	0.55	0.54	0.51	0.47
0.66	0.63	0.62	0.62	0.62	0.61	0.60	0.60	0.59	0.57	0.56	0.53	0.49
0.68	0.65	0.64	0.64	0.64	0.63	0.62	0.62	0.61	0.60	0.58	0.56	0.51
0.70	0.67	0.66	0.66	0.66	0.65	0.65	0.64	0.63	0.62	0.60	0.58	0.54
0.72	0.69	0.69	0.68	0.68	0.67	0.67	0.66	0.65	0.64	0.63	0.60	0.56
0.74	0.71	0.71	0.70	0.70	0.69	0.69	0.68	0.67	0.66	0.65	0.63	0.59
0.76	0.73	0.73	0.72	0.72	0.72	0.71	0.71	0.70	0.69	0.67	0.65	0.61
0.78	0.75	0.75	0.75	0.74	0.74	0.73	0.73	0.72	0.71	0.70	0.68	0.64
0.80	0.77	0.77	0.77	0.76	0.76	0.76	0.75	0.74	0.73	0.72	0.70	0.67
0.82	0.79	0.79	0.79	0.79	0.78	0.78	0.77	0.77	0.76	0.75	0.73	0.70
0.84	0.82	0.81	0.81	0.81	0.81	0.80	0.80	0.79	0.78	0.77	0.76	0.73
0.86	0.84	0.84	0.83	0.83	0.83	0.82	0.82	0.81	0.81	0.80	0.78	0.75
0.88	0.86	0.86	0.86	0.85	0.85	0.85	0.84	0.84	0.83	0.82	0.81	0.79
0.90	0.88	0.88	0.88	0.88	0.87	0.87	0.87	0.86	0.86	0.85	0.84	0.82

Tabela 6.1: *in* J. A. Hanley e B. J. McNeil, A Method of Comparing the Areas under Receiver Operating Characteristic Curves derived from the Sames Cases, *Radiology*, 148:839-843,1983.A primeira coluna é relativa à média dos coeficientes de correlação e a primeira linha é relativa à média das áreas

Parte II

Output programa SPSS: Curva ROC empírica

Ouput Programa SPSS: Curva ROC empírica

Ouput Programa SPSS: Curva ROC empírica

Ouput Programa SPSS: Curva ROC empírica

Ouput Programa SPSS: Curva ROC empírica

Parte III

Output programa ROCKIT:

Modelo binormal

Output programa ROCKIT: Modelo binormal

Output programa ROCKIT: Modelo binormal

Output programa ROCKIT: Modelo binormal

Output programa ROCKIT: Modelo binormal

Output programa ROCKIT: Modelo binormal

Output programa ROCKIT: Modelo binormal

Parte IV

Output PROPROC: Modelo binormal próprio

Output programa PROPROC: Modelo binormal próprio

Output programa PROPROC: Modelo binormal próprio

Output programa PROPROC: Modelo binormal próprio

Parte V

Subrotina Fortran: *kernel*

Gaussiana

Subrotina Fortran: *kernel* Gaussiana

Parte VI

Subrotina Fortran: *kernel*

biweight

Subrotina Fortran: *kernel* biweight

Parte VII

Programa Fortran: Curva ROC

kernel

Programa Fortran: Curva ROC *kernel*

Programa Fortran: Curva ROC *kernel*

Bibliografia

[1] Alavi. Observer performance methodology in medical imaging. Notes for Dr. Alavis's Course.

[2] T. A. Alonzo and M. S. Pepe. Distribution-free roc analysis using binary regression techniques. *Biostatistics*, 3(3):421–432, 2002.

[3] C. B. Begg. Advances in statistical methodology for diagnostic medicine in the 1980's. *Statistics in Medicine*, 10:1887–1895, 1991.

[4] D. A. Bloch. Comparing two diagnostic tests against the same "gold standard"in the same sample. *Biometrics*, 53:73–85, 1997.

[5] G. Campbell. General methodology 1 - advances in statistical methodology for the evaluation of diagnostic and laboratory tests. *Statistics in Medicine*, 13:499–508, 1994.

[6] A. V. Carneiro. Princípios de seleccção e uso de testes de diagnóstico: Propriedades intrínsecas dos testes. *Revista Portuguesa de Cardiologia*, (12):1267–1274, 2001.

[7] L. C.J. Using smoothed receiver operating characteristic curves to summarize and compare diagnostic systems. *Journal of the American Statistical Association*, 93:1356–1364, 1998.

[8] C. J. Clopper and E. Pearson. The use of confidence of fiducial limits ilustrated in the case of the binomial. *Biometrika*, 26:404, 1934.

[9] A. C. da Silva Braga. *Curvas ROC: Aspectos Funcionais e Aplicações*. PhD thesis, Universidade do Minho, 2000.

[10] Diamond. Roc steady, a receiver operating characteristic curve that is invariant relative to selection bias. 1987.

[11] D. E.R., D. D. M., and D. Clarke-Pearson. Comparing the areas under two or more correlated receiver operating characteristic curves: Nonparametric approach. *Biometrics*, 44:837–845, 1998.

[12] D. Faraggi and B. Reiser. Estimation of the area under the roc curve. *Statistics in Medicine*, 31:3093–3106, 2002.

[13] M. Goddard and I. Hinberg. Receiver operator characteristic (roc) curves and non-normal data: an empirical study. *Statistics in Medicine*, 9:325–337, 1990.

[14] D. J. Goodenough, K. Rossmann, and L. B. Lusted. Radiographic applications of signal detection theory. *Radiology*, (105):199–200, 1972.

[15] D. J. Goodenough, K. Rossmann, and L. B. Lusted. Radiographic applications of receiver operating characteristic (roc) curves. *Radiology*, (110):89–95, 1974.

[16] J. A. Hanley. The use os the binormal model for parametric roc analysis of quantitative diagnostic tests. *Statistics in Medicine*, 15:1575–1585, 1996.

[17] J. A. Hanley and B. J. McNeil. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143, 1982.

[18] J. A. Hanley and B. J. McNeil. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology*, (148):839–843, 1983.

[19] F. Hsieh and B. W. Turnbull. Nonparametric and semiparametric estimation of the receiver operating characteristic curve. *The Annals of Statistics*, 24:25–40, 1996.

[20] K. Jensen, H.-H. Muller, and H. Schafer. Regional conidence bands for roc curves. *Statistics in Medicine*, 19:493–509, 2000.

[21] Y. J. Jiang, C. E. Metz, and R. M. Nishikawa. A receiver operating characteristic partial area index for higly sensitive diagnostic tests. *Radiology*, (201):745–750, 1996.

[22] W. B. Langdon. The magnificent roc, 2001.

[23] P. A. W. Lewis and E. J. Oraw. *Simulation Methodology for Statisticians, Operations Analysts and Engineers.* Wadsworth, Inc., Belmont, California, 1989.

[24] C. J. Lloyd. Fitting roc curves using non-linear binomial regression. *Australian N. Z. J. Statistical*, 42(2):193–204, 2000.

[25] C. J. Lloyd. Semi-parametric estimation of roc curves based on binomial regression modelling. *Australian N. Z. J. Statistical*, 44(1):75–86, 2002.

[26] L. Lusted. *Introduction to Medical Decision Making.* C.C. Thomas, Springfield, 1968.

[27] F. A. Mann, C. F. Hildebolt, and A. J. Wilson. Statistical analysis with receiver operating characteristic curves. *Radiology*, (184):37–38, 1992.

[28] C. Metz and X. Pan. "proper"binormal roc curves: Theory and maximum-likelihood estimation. *Journal of Mathematical Psychology*, 43:1–33, 1999.

[29] C. E. Metz. Statistical analysis of roc data in evaluating diagnostic performance. Multiple Regression Analysis: Applications in the helth sicences, number 13, edited by Donald E. Herbert and Raymond H. Meyrs.365-384. American Intitute of Physics, 1986.

[30] C. E. Metz. *IBM compatable ROCKIT User's Guide - ROCKIT 0.9B Beta Version*, 1998.

[31] C. E. Metz, B. A. Herman, and C. A. Roe. Statistical comparison of two roc-curve estimates obtained from partially-paired datasets. *Medical Decisiom Making*, 18:110–121, 1998.

[32] H. B. Metz C.E. and S. J-H. Maximum likelihood estimation of receiver operating characteristic (roc) curves from continuously-distributed data. *Statistics in Medicine*, 17:1033–1053, 1998.

[33] L. E. Moses, D. Shapiro, and B. Littenberg. Combining independent studies of a diagnostic test into a summary roc curve: Data-analytic approaches and some additional considerations. *Statistics in Medicine*, 12:1293–1316, 1993.

[34] D. Mossman. Three-way rocs. *Medical Decision Making*, (19):78–89, 1999.

[35] D. D. Pestana and S. F. Velosa. *Introdução à Probabilidade e à Estatística*, volume 1. Fundação Calouste Gulbenkian, 2002.

[36] H. Schafer. Constructing a cut-off point for a quantitative diagnostic test. *Statistics in Medicine*, 8:1381–1391, 1989.

[37] H. Schafer. Efficient confidence bounds for roc curves. *Statistics in Medicine*, 13:1551–1561, 1994.

[38] A. Sorribas, J. March, and J. Trujillano. A new parametric method based on s-distributions for computing receiver operating characteristics curves for continuous tests. *Statistics in Medicine*, 21:1213–1235, 2002.

[39] S. M. Spalding, V. Wald, and L. Bernd. Ige sérica total em atópicos e não atópicos na cidade de porto alegre. *Associação Médica do Brasil*, 46(2):93–7, 2000.

[40] J. Swets and R. Pickett. *Evaluation of Diagnostic Systems: Methods from Signal Detection Theory*. Academic Press, New York.

[41] M. L. Thompson and W. Zucchini. On the statistical analysis of roc curves. *Statistics in Medicine*, 8:1277–1290, 1989.

[42] Tosteson and Begg. A general regression methodology for roc curve estimation. 1988.

[43] S. D. Walter. Properties of the summary receiver operating characteristic (sroc) curve for diagnostic test data. *Statistics in Medicine*, 21:1237–1256, 2002.

[44] J. W.Shaw and W. Horrace. Comparison of nonparametric receiver operating characteristic analysis with a likelihood-ratio test for model selection.

[45] H. Yang and D. Carlin. Roc surface: A generalization of roc curve analysis. *Journal of Biopharmaceutical Statistics*, (10):183–196, 2000.

[46] D. D. Zhang, X.-H. Zhou, D. H. F. Jr., and L. Freeman. A non-parametric method for the comparison for partial areas under roc curves and its application to large health care data sets. *Statistics in Medicine*, 21:701–715, 2002.

[47] X. H. Zhou and C. A. Gatsonis. A simple method for comparing correlated roc curves using incomplete data. *Statistics in Medicine*, 15:1687–1693, 1996.

[48] K. H. Zou, W. Hall, and D. E. Shapiro. smooth non-parametric receiver operating characteristic (roc) curves for continuous diagnostic tests. *Statistics in Medicine*, 16:2143–2156, 1997.