Regiões de Incerteza para a Curva ROC em Testes Diagnósticos

Janaina Cândida Lopes Vaz

Regiões de Incerteza para a Curva *ROC* em Testes Diagnósticos

Janaina Cândida Lopes Vaz

Orientador: Prof. Dr. Luis Aparecido Milan

Dissertação apresentada ao Departamento de Estatística da Universidade Federal de São Carlos - DEs/UFSCar, como parte dos requisitos para obtenção do título de Mestre em Estatística.

UFSCar - São Carlos

Março / 2009

Ficha catalográfica elaborada pelo DePT da Biblioteca Comunitária da UFSCar

V393ri

Vaz, Janaina Cândida Lopes.
Regiões de incerteza para a curva ROC em testes diagnósticos / Janaina Cândida Lopes Vaz. -- São Carlos : UFSCar, 2009.
151 f.

Dissertação (Mestrado) -- Universidade Federal de São Carlos, 2009.

1. Estatística matemática. 2. Inferência clássica. 3. Estatística médica. 4. Análise de regressão. I. Título.

CDD: 519.5 (20ª)

UNIVERSIDADE FEDERAL DE SÃO CARLOS
Centro de Ciências Exatas e de Tecnologia
Programa de Pós-Graduação em Estatística
Via Washington Luís, Km 235 - C.P.676 - CGC 45358058/0001-40
FONE: (016) 3351-8292/3351-8241 - FAX: (016) 3351-8243
13565-905 - SÃO CARLOS-SP-BRASIL
www.ufscar.br/~des ppgest@power.ufscar.br

Pós-Graduação UFSCar
multiplicando conhecimento

DECLARAÇÃO

Declaramos, para os devidos fins, que Janaina Cândida Lopes Vaz defendeu sua Dissertação de Mestrado no dia 03/03/2009, tendo sido **aprovada**. A aluna deverá apresentar a versão final da dissertação (com as correções e sugestões da Banca, e a ficha catalográfica anexada), e a Certidão Negativa da Biblioteca Comunitária, para formação do processo de homologação e emissão do Diploma do Título.

Igualmente, a aluna deverá apresentar a documentação da pesquisa (rotinas, arquivos em LaTeX, resultados complementares etc.) ao seu orientador, visando facilitar a confecção de relatórios técnicos que condensarão os resultados obtidos.

Essa declaração é válida pelo período de 30 dias.

UNIVERSIDADE FEDERAL PPG-Es SÃO CARLOS

Prof. Dr. Josemar Rodrigues
Coordenador – PPG-Es / UFSCar

Para meus pais Domingos e Ercília

Agradecimentos

A Deus, pela vida, por seu amor, pela proteção, por ter me dado força e oportunidade para chegar até aqui.

Aos meus pais Domingos e Ercília, por ser o meu porto seguro e meu refúgio para todas as horas. A eles, meu muito obrigada pelo amor infinito e único, pela dedicação de uma vida inteira, me dando sempre muito carinho, me mostrando desde cedo a importância da educação e me ajudando em todos os sentidos: financeiramente, espiritualmente, fraternalmente. Pais, vocês são minha vida.

À minha irmã Jamile (carinhosamente chamada de "Mil"), pelo amor, pela amizade, pelo companheirismo e pelo incentivo em todas minhas decisões. Mil, amo-te incondicionalmente.

Ao Professor Milan, pelo incentivo, pelo carinho, pela disponibilidade, pela dedicação e por toda paciência, com que sempre me recebeu e orientou durante esse Mestrado.

Aos meus amigos: de infância, da Fundação, da UNESP, da UFSCar, de Tanabi, de São José do Rio Preto, de São Carlos e de São Paulo, pela amizade e pelos diversos momentos que passamos juntos.

A todos meus familiares, em especial, aos meus primos, pelo companheirismo, pela profunda amizade e pelos vários momentos que compartilhamos.

À minha avó Dirce (*in memoriam*), pelas orações feitas e pelo carinho.

Ao Edson Z. Martinez, ao Sofus A. Macskassy, ao Foster Provost e ao Pedro Emmanuel A. do Brasil, que me ajudaram enviando artigos solicitados.

Ao Saulo Morellato, ao Leandro Lopes Teixeira e, em especial, ao Luis Ernesto Bueno Salasar, que me ajudaram com parte da programação.

Aos professores Francisco Louzada Neto e Mariana Cúri, membros da banca de defesa da Dissertação, e aos professores Teresa Cristina Martins Dias e Francisco Louzada Neto, membros da banca do exame de Qualificação, pelas observações e sugestões.

Aos professores Francisco Louzada Neto e Mariana Cúri, membros da banca de defesa da Dissertação, pelos comentários e sugestões.

A todos do Departamento de Estatística da Universidade Federal de São Carlos, em especial, à professora Maria Sílvia de Assis Moura, ao professor Carlos Alberto Ribeiro Diniz, à "Tia Luiza" e à Isabel, que me ajudaram e contribuíram com meu amadurecimento pesssoal e profissional, sempre com palavras oportunas.

À empresa Refrigerantes Arco Íris Ltda, em especial, ao meu tio Erasmo Antônio Lopes Perez e ao Fabiano Oliveira de Almeida, que me ajudaram com o transporte de Tanabi para São Carlos, e vice-versa, durante todo o Mestrado.

À Coordenação de Aperfeiçoamento Pessoal de Nível Superior (CAPES) pelo apoio financeiro.

A todos os demais, que de alguma forma, contribuíram para a finalização deste trabalho.

Resumo

Testes diagnósticos são métodos capazes de indicar a presença ou ausência de uma doença, com uma probabilidade de erro. O desempenho de um teste diagnóstico pode ser verificado por algum indicador, como: a especificidade, a sensibilidade e a curva *ROC*. Um gráfico do complemento da especificidade *versus* sensibilidade é chamado de curva *ROC*. A curva *ROC* demonstra a habilidade do teste em discriminar os diferentes diagnósticos da doença, logo é uma ferramenta gráfica que serve para avaliar o desempenho de um teste. Definimos três tipos de regiões de confiança em torno da curva *ROC*: as pontuais, as regionais e as globais. Em algumas situações, de acordo com a necessidade do clínico, uma decisão é tomada sobre uma determinada região específica da curva *ROC*. Revisamos alguns procedimentos para estimar a região de confiança para a curva *ROC* e propomos dois novos métodos (médias otimizadas e médias limiares otimizadas) para estimar essa região. Usamos o método *bootstrap* para buscar uma região de confiança em torno da curva *ROC*. Usando exemplos numéricos, aplicamos os métodos para comparar seus desempenhos.

Palavras-chave: Teste Diagnóstico; Curva *ROC*; *Bootstrap*; Regiões de Confiança para Curva *ROC*.

Abstract

Diagnostic tests are methods capable of indicating the presence or absence of a disease, with a probability of error. The performance of a diagnostic test can be verified by some indicator, as: the specificity, the sensitivity and the ROC curve. A graph of the specificity complement versus sensitivity is called as ROC curve. The ROC curve demonstrates the test's ability to discriminate the different disease diagnosis, therefore it is a graphical tool that is used to assess the performance of a test. We define three types of confidence regions around the ROC curve: the punctual, the regional and the global. In some instances, depending on the clinical needs, the decision is taken under an specific region of the ROC curve. We review some procedures for estimating confidence region for the ROC curve and we propose two new methods (optimized averages and averages thresholds optimized) to estimating that region. We use the bootstrap method to search for a confidence region around the ROC curve. Using numerical examples, we apply the methods an compare their performance.

Keywords: Diagnostic Tests; ROC Curve; Bootstrap; Confidence Bands for ROC Curves.

Lista de Abreviaturas

B: Número de reamostragens *bootstrap*

D: Condição do paciente

$Dist_q$: Distância máxima vertical entre $F(h)$ e $F_q(h)$

E_S: Especificidade

FN: Falso negativo

FP: Falso positivo

FWB: Bandas de largura fixa

$F(h)$: Função de distribuição acumulada dos dados

$F_q(h)$: Função de distribuição acumulada empírica dos dados

g: Distância horizontal ao longo dos valores de FP

G: Variável de decisão

h: Distância vertical ao longo dos valores de VP

i: Quantidade de FP usada para geração dos pontos de corte empíricos

I: Função indicadora

j: Número de curvas ROC geradas durante o *bootstrap*

KS: *Kolmogorov-Smirnov*

l: Comprimento do intervalo gerado no eixo das abscissas

m: Número total de indivíduos doentes

MLO: Médias limiares otimizadas

MO: Médias otimizadas

n: Número total de indivíduos não doentes

R: Resultado do teste diagnóstico

ROC: *Receiver operating characteristic*

r_0: Ponto de corte

S: Variável aleatória contínua que representa o resultado do teste diagnóstico

S_E: Sensibilidade

SJR: Região de confiança de juntas simultâneas

t: Número de trapézios sob a curva ROC

TA: Médias limiares

$t_{(\alpha)}$: 100α-ésimo percentil da distribuição t-Student

VA: Médias verticais

Var: Variância

VN: Verdadeiro negativo

VP: Verdadeiro positivo

VPN: Valor predito negativo

VPP: Valor predito positivo

X: Resultado do teste diagnóstico para indivíduos não doentes

Y: Resultado do teste diagnóstico para indivíduos doentes

$z_{(\alpha)}$: 100α-ésimo percentil da distribuição normal

α: Nível de significância

β: Viés

θ: Quantidade populacional

π: Prevalência

ϱ: Inclinação para gerar bandas de largura fixa

τ: Probabilidade do indivíduo ser registrado como doente

Sumário

Capítulo 1

Introdução

1.1 Motivação

Qualquer pesquisador, deparado com a necessidade de analisar determinados dados, precisa fazer uma escolha sobre o método investigativo a ser utilizado nessa análise. É necessário que algumas considerações importantes sejam feitas nessa escolha, como por exemplo:

- O objetivo da investigação;

- As características matemáticas das variáveis envolvidas;

- As hipóteses estatísticas feitas sobre estas variáveis;

- Como os dados foram coletados.

Para certos acontecimentos, existem testes baseados quer em observações de um determinado fenômeno, quer em técnicas laboratoriais, que permitem a previsão e detecção desses acontecimentos numa fase inicial do seu desenvolvimento.

Um estímulo para o desenvolvimento desse estudo é o problema da discriminação existente em testes diagnósticos, que consiste em classificar de uma forma precisa, os casos considerados normais (indivíduos não doentes) e os anormais (indivíduos doentes).

Em testes diagnósticos é necessário que conheçamos sobre a exatidão

e a precisão dos mesmos. A precisão está associada à dispersão dos valores em sucessivas observações, enquanto que a exatidão refere-se ao quão próxima uma estimativa deve estar do verdadeiro valor que ela pretente representar. As limitações da exatidão e da precisão durante o diagnóstico originam a introdução dos conceitos de sensibilidade e especificidade de um teste diagnóstico.

Segundo Metz (1986), sensibilidade é a probabilidade de decidir se a doença em questão está presente quando, de fato, está presente, e especificidade é a probabilidade de decidir se a doença em questão está ausente quando, de fato, está ausente. Estas medidas e os índices a elas associados, como por exemplo, a proporção de verdadeiros positivos e a proporção de falsos positivos, são mais significantes do que a exatidão, embora não forneçam uma descrição única do desempenho do diagnóstico.

O maior problema sobre sensibilidade e especificidade é que essas medidas dependem do critério do diagnóstico ou de um valor de corte, o qual é, por vezes, selecionado arbitrariamente. Assim, quando mudado o critério, pode ocorrer um aumento na sensibilidade e consequentemente uma diminuição na especificidade, e vice-versa. Logo, elas representam um quadro incompleto do desempenho de um teste diagnóstico.

Um critério de decisão particular depende também dos benifícios associados aos resultados corretos e dos custos associados aos incorretos. Por exemplo, a previsão de uma tempestade que acaba por não ocorrer (falso positivo) é tipicamente vista como tendo um custo bem menor do que em relação a uma falha na previsão de uma tempestade que ocorre (falso negativo). Logo, o critério adotado para um diagnóstico negativo deve ser mais cauteloso em relação àquele adotado para um diagnóstico positivo.

Existem dois erros que podem ocorrer durante a decisão do diagnóstico de um teste: a escolha de uma falha (no sentido de declarar um doente como um indivíduo não doente) ou a escolha de um alarme falso (declarar um indivíduo não doente como doente). Por exemplo, para um profissional que tem diante de si o diagnóstico de uma determinada doença, ao ter de se decidir, ele irá preferir um alarme falso a uma falha. Assim, esse profissional irá optar por um teste

mais sensível. Por outro lado, ele deve sempre estar consciente que uma terapia disponível para este tipo de doença pode ser cara e deficiente, o que torna o teste pouco específico.

Para contornar esse tipo de situação, foi necessário desenvolver medidas alternativas de diagnóstico com propriedades mais robustas do que a sensibilidade e a especificidade. Assim, surgiu a curva *ROC* (um acrônimo em inglês - *receiver operating characteristics*). A análise *ROC* pode ser efetuada através de um método gráfico (curva *ROC*) e o desempenho de um dado teste pode ser avaliado através de índices de precisão associados à curva *ROC*, como por exemplo, a área abaixo desta ou regiões de confiança em torno da curva.

1.2 Estrutura da dissertação

O conjunto de objetivos propostos na seção anterior representam, ainda que parcialmente, o modo como o trabalho foi estruturado.

No Capítulo 2, apresentamos os conceitos principais envolvidos na análise de um teste diagnóstico, mediante comparação com um teste padrão-ouro, tais como: prevalência e seu complemento, valores predito (positivo e negativo), condições de acerto (verdadeiro positivo e verdadeiro negativo), condições de erro (falso positivo e falso negativo) e medidas de desempenho (sensibilidade, especificidade e seus respectivos complementos). Também encontramos a função de verossimilhança para o teste investigado e o estimador de máxima verossimilhança para a sensibilidade, para a especificidade e para a prevalência.

No Capítulo 3, baseado na teoria de testes diagnósticos, descrevemos uma curva *ROC*. Neste, relatamos sobre a medida de precisão da curva *ROC* (área sob a curva). Ao longo do mesmo, também tratamos sobre a perspectiva histórica que envolve a curva *ROC*, bem como uma relação da mesma com a teoria estatística e a teoria da detecção do sinal. Finalmente, construímos uma curva *ROC* para cada um dos dois testes diagnósticos utilizados na Aplicação e encontramos a área sob essas curvas, a fim de investigar qual dos dois testes tem melhor desempenho.

No Capítulo 4, enunciamos o processo de simulação via reamostragem *bootstrap*, tanto sua forma paramétrica quanto a não paramétrica. Também tratamos sobre intervalos de confiança *bootstrap* percentil, normal, studentizado e binomial, e apresentamos um método para geração de uma região de incerteza *bootstrap* para a curva *ROC*.

O Capítulo 5 é compreendido de toda uma teoria sobre a região de confiança, ora chamada de região de incerteza, ora de região de probabilidade, associada à curva *ROC*. Neste, tratamos de regiões de confiança geradas de formas distintas, cada qual relacionada a uma necessidade diferente. Assim, apresentamos uma teoria sobre geração de bandas pontuais (métodos das: médias verticais, médias otimizadas, médias limiares e médias limiares otimizadas), bandas regionais (método das regiões de confiança de juntas simultâneas) e bandas globais (método das bandas de largura fixa). Também mostramos de que maneira essas bandas são geradas.

No Capítulo 6, são apresentadas algumas aplicações recorrendo à análise de dados, para um específico teste diagnóstico, através da curva *ROC*. Calculamos intervalos de confiança e regiões de incerteza em torno dessa curva. Para tal, é utilizado um conjunto de dados reais, apresentado por Zhou *et al.* (2002), que diz respeito a um exame de mamografia realizado em 60 mulheres.

Finalmente, apresentamos as conclusões gerais sobre o trabalho realizado e apontamos algumas linhas de orientação para pesquisas futuras.

Capítulo 2

Testes Diagnósticos

Nas últimas décadas, métodos estatísticos aplicados à medicina diagnóstica sofreram grandes avanços. A maior parte destes está voltada para questão de classificar indivíduos em grupos, sendo que testes diagnósticos compõem o principal exemplo.

O procedimento diagnóstico é um processo que inclui um razoável grau de incerteza, que é elevado ou diminuído na dependência de um bom juízo por parte dos médicos, além de um vasto conhecimento da literatura médica. Assim, a prática médica moderna utiliza das leis da probabilidade como um importante auxiliar na interpretação de testes diagnósticos (Saraiva, 2004).

Para estar certo se a doença está realmente presente ou ausente em determinado paciente, devemos lançar mão de testes relativamente elaborados, caros ou arriscados. Dentre esses estão a biópsia, a exploração cirúrgica e, certamente, a autópsia. Como os meios mais acurados de estabelecer um diagnóstico correto são, quase sempre, mais caros e invasivos, clínicos e pacientes preferem testes mais simples do que um teste padrão rigoroso, pelo menos inicialmente. Um exemplo disso é que eletrocardiogramas e enzimas séricas são habitualmente usados para estabelecer o diagnóstico de infarto agudo do miocárdio, em vez de cateterismo ou exames de imagem. Entretanto, o uso de testes mais simples, como substitutos dos mais elaborados e exatos, resulta num risco maior de um diagnóstico incorreto.

Testes diagnósticos são descritos como métodos capazes de indicar a

presença ou ausência de uma doença, com uma determinada chance de erro (Linnet, 1987). Esses erros são principalmente de duas formas: classificar um indivíduo doente como não doente e classificar um indivíduo não doente como doente.

Os estudos de avaliação de testes diagnósticos comparam um teste clínico ou laboratorial em desenvolvimento com um padrão-ouro (*gold-standard*, também conhecido como teste de referência), isto é, um procedimento que deseja identificar a condição de doente ou não doente de um determinado indivíduo (Riegelman e Hirsch, 1996). Entretanto, o padrão-ouro é frequentemente difícil de ser encontrado, pois ele é resultante de um teste que pode ser doloroso, caro, invasivo, demorado e até mesmo inviável. Assim, ele não é aplicado corriqueiramente (Sox, 1986) e nem sempre é desprovido de erros, podendo ocasionar problemas na busca de medidas de precisão para o teste sob investigação (Thibodeau, 1981).

É importante que o clínico procure um padrão-ouro que descreva o estado do indivíduo (doente ou não doente) com maior veracidade possível, pois, muitas vezes, ele pode oferecer riscos aos pacientes.

Em cada indivíduo, a determinação do resultado do teste sob investigação não deve ser influenciada pelo conhecimento prévio do resultado do padrão-ouro, o que fornecerá dados viciados.

Quando o padrão-ouro é comparado a um teste sob investigação, ambos fornecem uma resposta dicotômica (positivo ou negativo), e observamos quatro situações relacionadas aos tipos de acerto e de erro.

Condições de acerto:

- ✓ Verdadeiro positivo (VP) - Probabilidade de um indivíduo doente ser classificado como doente;
- ✓ Verdadeiro negativo (VN) - Probabilidade de um indivíduo não doente ser classificado como não doente.

Condições de erro:

- ✓ Falso positivo (FP) - Probabilidade de um indivíduo não doente ser classificado como doente;
- ✓ Falso negativo (FN) - Probabilidade de um indivíduo doente ser classificado como não doente.

Os valores de VP e VN, como vistos anteriormente, são interpretados como "acerto" do teste sob investigação, e são importantes medidas de desempenho. Dentre as situações interpretadas como "erro" do teste, os valores de FN merecem uma atenção maior que os valores de FP. Se o resultado do teste sob investigação é negativo, mas o indivíduo é portador da doença, o fato dele acreditar no teste vai fazer com que ele não trate sua enfermidade, ocasionando, muitas vezes, consequências irreparáveis para o mesmo. Um indivíduo portador dos sintomas de uma determinada doença, mas com um resultado negativo do teste diagnóstico, é um candidato a se enquadrar na situação de FN. As consequências de um resultado FP também poderão ser maléficas ao indivíduo.

No entanto, um veredito final a respeito do resultado de um teste deverá estar acompanhado do conhecimento e da experiência do profissional que investiga a doença que o teste pretende predizer.

Para conceituar as medidas descritas, considere um conjunto de dados que mostra a relação entre os resultados de dois testes diagnósticos, sendo um deles considerado padrão-ouro. A Figura 2.1 mostra a quantidade de pacientes classificados como positivos ou negativos pelo teste diagnóstico que está sendo investigado, e também, dentre estes, quantos foram identificados pelo teste padrão-ouro como doentes (portadores da doença) ou não doentes (não portadores da doença).

		PADRÃO-OURO		
		Doente	**Não doente**	**Total**
TESTE	**Positivo**	VP	FP	VP + FP
	Negativo	FN	VN	FN +VN
	Total	VP + FN	FP + VN	VP + FP + FN +VN

FIGURA 2.1: Possíveis resultados de um teste diagnóstico para identificar uma doença.

2.1 Medidas de desempenho

No estudo de testes diagnósticos, um procedimento importante é a obtenção das medidas de desempenho do teste sob investigação. Valores preditos, sensibilidade, especificidade, bem como outros elementos estatísticos da teoria da decisão, são medidas de desempenho usadas para avaliar testes diagnósticos (Fleiss, 1981).

Considere as variáveis aleatórias discretas D e R, como sendo correspondentes, respectivamente, à verdadeira condição do paciente (o qual é uma realidade desconhecida e que provém do padrão-ouro) e o resultado do teste diagnóstico. Cada uma dessas variáveis aleatórias assumem somente os valores 0 ou 1. As interpretações para elas são:

$$\begin{cases} \mathrm{D} = 1, & \text{se o paciente é doente,} \\ \mathrm{D} = 0, & \text{se o paciente não é doente,} \end{cases}$$

e

$$\begin{cases} \mathrm{R} = 1, & \text{se o resultado do teste é positivo,} \\ \mathrm{R} = 0, & \text{se o resultado do teste é negativo.} \end{cases}$$

A **prevalência** (π) de uma doença é definida como a proporção de indivíduos portadores da doença que está sendo investigada em uma determinada

população, durante um determinado intervalo de tempo. A prevalência pode ser interpretada como a probabilidade de um indivíduo ser portador da doença sob investigação, antes mesmo dele ser submetido a um teste diagnóstico. Logo, a prevalência é a probabilidade a *priori*. Se as informações são registradas com precisão em um teste diagnóstico, então os dados dicotômicos seguem um processo de Bernoulli com parâmetro π. Assim, a prevalência expressa a probabilidade da doença antes mesmo do teste ser realizado, por isso também é denominada de probabilidade prévia ou pré-teste. O valor de π pode ser estimado da seguinte forma

$$\widehat{\pi} = \frac{VP + FN}{VP + FP + FN + VN}.$$

Em termos de probabilidade, a prevalência é dada por

$$\pi = P(D = 1).$$

Complemento da prevalência $(1 - \pi)$ é definido como a probabilidade de um indivíduo não ser portador da doença sob investigação, antes mesmo dele ser submetido a um teste diagnóstico. Em outras palavras, $1-\pi$ refere-se à todos os casos da doença inexistentes previamente à realização de um teste diagnóstico. O valor de $1 - \pi$ pode ser estimado da seguinte forma

$$\widehat{\pi}^c = 1 - \widehat{\pi} = \frac{FP + VN}{VP + FP + FN + VN}.$$

Em termos de probabilidade, o complemento da prevalência é dado por

$$\pi^c = 1 - \pi = P(D = 0).$$

Valor predito positivo (VPP) de um teste diagnóstico é definido como a proporção de indivíduos portadores da doença, dado que o resultado do teste é positivo. Em outras palavras, VPP expressa a probabilidade de um paciente com resultado positivo no teste ter a doença. Segundo Dunn e Everitt (1995),

este valor pode ser estimado diretamente da amostra da seguinte forma

$$\widehat{VPP} = \frac{VP}{VP + FP}.$$

Em termos de probabilidade, o valor predito positivo é dado por

$$VPP = P(D = 1 \mid R = 1).$$

Valor predito negativo (VPN) de um teste diagnóstico é definido como a proporção de indivíduos não portadores da doença, dado que o resultado do teste é negativo. Em outras palavras, VPN expressa a probabilidade de um paciente com resultado negativo no teste não ter a doença. Segundo Dunn e Everitt (1995), este valor pode ser estimado diretamente da amostra da seguinte forma

$$\widehat{VPN} = \frac{VN}{FN + VN}.$$

Em termos de probabilidade, o valor predito negativo é dado por

$$VPN = P(D = 0 \mid R = 0).$$

A interpretação de VPP e VPN deve ser cautelosa, pois estas medidas sofrem efeito da prevalência da doença. Se a amostra utilizada para investigar o desempenho do teste é obtida através de um delineamento experimental que não permite um estimador de prevalência, os valores preditos podem não ser medidas confiáveis. Por esse motivo, as medidas de desempenho mais utilizadas são sensibilidade e especificidade, uma vez que elas não são afetadas pela prevalência da doença (Altman, 1991).

Sensibilidade (S_E) é definida como a probabilidade do teste que está sendo investigado fornecer um resultado positivo, dado que o indivíduo é portador da doença. Sensibilidade também é conhecida como taxa de VP (TVP) ou taxa $1 - FN$, ou seja, é a proporção de VP entre todos os doentes. A quantidade S_E

avalia a capacidade do teste detectar a doença quando ela está de fato presente. O valor de S_E pode ser estimado da seguinte forma

$$\widehat{S_E} = \frac{VP}{VP + FN}.$$

Em termos de probabilidade, a sensibilidade é dada por

$$S_E = P(R = 1 \mid D = 1).$$

Complemento da sensibilidade $(1 - S_E)$ é definido como a probabilidade do teste que está sendo investigado fornecer um resultado negativo, dado que o indivíduo é portador da doença. Complemento da sensibilidade também é conhecido como taxa de FN (TFN) ou taxa de $1 - VP$, ou seja, é a proporção de FN entre todos os doentes. A quantidade $1 - S_E$ avalia a capacidade do teste não detectar a doença quando ela está de fato presente. O valor de $1 - S_E$ pode ser estimado da seguinte forma

$$\widehat{S_E}^c = 1 - \widehat{S_E} = \frac{FN}{VP + FN}.$$

Em termos de probabilidade, o complemento da sensibilidade é dado por

$$S_E^c = 1 - S_E = P(R = 0 \mid D = 1).$$

Especificidade (E_S) é definida como a probabilidade do teste que está sendo investigado fornecer um resultado negativo, dado que o indivíduo não é portador da doença. Especificidade também é conhecida como taxa de VN (TVN) ou taxa de $1 - FP$, ou seja, é a proporção de VN entre todos os não doentes. A quantidade E_S avalia a capacidade do teste "afastar" a doença quando ela está de fato ausente. O valor de E_S pode ser estimado da seguinte forma

$$\widehat{E_S} = \frac{VN}{FP + VN}.$$

Em termos de probabilidade, a especificidade é dada por

$$E_S = P(R = 0 \mid D = 0).$$

Complemento da especificidade $(1 - E_S)$ é definido como a probabilidade do teste que está sendo investigado fornecer um resultado positivo, dado que o indivíduo não é portador da doença. Complemento da especificidade também é conhecido como taxa de FP (TFP) ou taxa de $1 - VN$, ou seja, é a proporção de FP entre todos os não doentes. A quantidade $1 - E_S$ avalia a capacidade do teste diagnosticar a doença quando ela está de fato ausente. O valor de $1 - E_S$ pode ser estimado da seguinte forma

$$\widehat{E_S}^c = 1 - \widehat{E_S} = \frac{FP}{FP + VN}.$$

Em termos de probabilidade, o complemento da especificidade é dado por

$$E_S^c = 1 - E_S = P(R = 1 \mid D = 0).$$

Como S_E e E_S não são calculadas sobre os mesmos indivíduos, temos que essas medidas são independentes entre si. A proporção de indivíduos doentes observada em estudos sobre desempenho de testes diagnósticos não interfere no cálculo de S_E e E_S, podendo assim afirmar que S_E e E_S não dependem da prevalência da doença.

Quando as estimativas de S_E e E_S são livres de viés (variabilidade entre subgrupos, variações entre observadores, viés devido à verificação), Linnet (1988) sugere que VPP e VPN podem ser estimados por

$$\widehat{VPP} = \frac{S_E\pi}{S_E\pi + (1 - E_S)(1 - \pi)},$$

e

$$\widehat{VPN} = \frac{E_S(1 - \pi)}{E_S(1 - \pi) + (1 - S_E)\pi}.$$

A **probabilidade de um teste diagnóstico ser positivo** (τ) é a probabilidade do indivíduo ser registrado como tendo a doença. Sendo D e R variáveis aleatórias independentes e considerando as leis de probabilidade condicional, temos que τ é dada por

$$\begin{aligned}\tau = P(R=1) &= P(D=1\ ,\ R=1)\ +\ P(D=0\ ,\ R=1)\\ &= P(R=1\ \cap\ D=1)\ +\ P(R=1\ \cap\ D=0)\\ &= P(R=1\mid D=1)\ P(D=1)\ +\ P(R=1\mid D=0)\ P(D=0)\\ &= S_E\ \pi\ +\ (1-E_S)(1-\pi) = S_E\ \pi\ +\ E_S^c\ \pi^c.\end{aligned}$$

A **probabilidade de um teste diagnóstico ser negativo** ($1\ -\ \tau$) é a probabilidade do indivíduo ser registrado como não tendo a doença. Sendo D e R variáveis aleatórias independentes e considerando as leis de probabilidade condicional, temos que $1-\tau$ é dada por

$$\begin{aligned}\tau^c = 1-\tau = P(R=0) &= P(D=1\ ,\ R=0)\ +\ P(D=0\ ,\ R=0)\\ &= P(R=0\ \cap\ D=1)\ +\ P(R=0\ \cap\ D=0)\\ &= P(R=0\mid D=1)\ P(D=1)\ +\ P(R=0\mid D=0)\ P(D=0)\\ &= (1-S_E)\ \pi\ +\ E_S\ (1-\pi) = S_E^c\ \pi\ +\ E_S\ \pi^c.\end{aligned}$$

Em ambos os casos citados, o registro de indivíduos doentes e não doentes segue um processo de Bernoulli com parâmetro τ ao invés de π.

As probabilidades $P(R=1)$ e $P(R=0)$ são chamadas de distribuições marginais da variável aleatória R.

2.2 Função de verossimilhança e estimador de máxima verossimilhança

Para encontrarmos a função de verossimilhança para o teste em questão, devemos considerar uma amostra aleatória com $n+m$ indivíduos selecionados

um de cada vez, com reposição, onde m representa o número total de indivíduos doentes e n representa o número total de indivíduos não doentes. Assim, a função de verossimilhança correspondente é proporcional a distribuição binomial com parâmetro τ, que é dada por

$$\begin{aligned} L(\pi, S_E, E_S \mid m+n, m, n) &\propto \tau^m \ (1-\tau)^n \qquad (2.1) \\ &\propto [S_E\pi + (1-E_S)(1-\pi)]^m \ [(1-S_E)\pi + E_S(1-\pi)]^n. \end{aligned}$$

Para encontrarmos o estimador de máxima verossimilhança dos parâmetros de interesse S_E, E_S e π, devemos encontrar cada um dos estimadores individualmente, supondo fixos os outros.

2.2.1 Estimador de máxima verossimilhança da S_E

Segundo Saraiva (2004), o EMV para o parâmetro S_E é obtido através do seguinte resultado:

Resultado 2.1 - O estimador de máxima verossimilhança da sensibilidade, quando os parâmetros E_S e π são fixos, é dado por

$$\widehat{S_E} = \begin{cases} 0, & \text{se } 1 + E_S\left(\frac{1}{\pi} - 1\right) - \frac{n}{(m+n)\pi} < 0, \\ 1 + E_S\left(\frac{1}{\pi} - 1\right) - \frac{n}{(m+n)\pi}, & \text{se } 0 \leq 1 + E_S\left(\frac{1}{\pi} - 1\right) - \frac{n}{(m+n)\pi} \leq 1, \\ 1, & \text{se } 1 + E_S\left(\frac{1}{\pi} - 1\right) - \frac{n}{(m+n)\pi} > 1. \end{cases}$$

Demonstração: Aplicando o logaritmo na função de verossimilhança 2.1, temos

$$\begin{aligned} l(\pi, S_E, E_S | m+n, m, n) &= \ln \left[\ (S_E\pi + (1-E_S)(1-\pi))^m \ ((1-S_E)\pi + E_S(1-\pi))^n \ \right] \\ &= m \ln(S_E\pi + (1-E_S)(1-\pi)) + n \ln((1-S_E)\pi + E_S(1-\pi)). \end{aligned}$$

Supondo fixos os parâmetros E_S e π, derivamos a função acima em relação a S_E, igualamos a mesma a zero e encontramos o EMV de S_E, ou seja,

$$\frac{\partial l(\pi, S_E, E_S|m+n, m, n)}{\partial S_E} = 0$$

$$\frac{m\pi}{S_E\pi + (1-E_S)(1-\pi)} + \frac{-n\pi}{(1-S_E)\pi + E_S(1-\pi)} = 0$$

$$\Rightarrow m\pi(\pi - S_E\pi + E_S - E_S\pi) = n\pi(S_E\pi + 1 - \pi - E_S + E_S\pi)$$

$$\Rightarrow m(\pi - S_E\pi + E_S - E_S\pi) = n(S_E\pi + 1 - \pi - E_S + E_S\pi)$$

$$\Rightarrow m\pi - m\pi S_E + mE_S - m\pi E_S = n\pi S_E + n - n\pi - nE_S + n\pi E_S$$

$$\Rightarrow m\pi + mE_S - m\pi E_S - n + n\pi + nE_S - n\pi E_S = n\pi S_E + m\pi S_E$$

$$\Rightarrow \pi(m+n) + E_S(m+n) - \pi E_S(m+n) - n = \pi S_E(m+n)$$

$$\Rightarrow (m+n)(\pi + E_S - \pi E_S) - \frac{n(m+n)}{m+n} = \pi S_E(m+n)$$

$$\Rightarrow \pi + E_S - \pi E_S - \frac{n}{m+n} = \pi S_E$$

$$\Rightarrow \widehat{S_E} = \frac{\pi + E_S - \pi E_S - \left(\frac{n}{m+n}\right)}{\pi} = 1 + \frac{E_S}{\pi} - E_S - \frac{n}{(m+n)\pi}$$

$$= 1 + E_S(\pi^{-1} - 1) - \frac{n}{(m+n)\pi}.$$

Logo,

$$\widehat{S_E} = 1 + E_S\left(\frac{1}{\pi} - 1\right) - \frac{n}{(m+n)\pi}. \qquad \blacksquare$$

2.2.2 Estimador de máxima verossimilhança da E_S

Segundo Saraiva (2004), o EMV para o parâmetro E_S é obtido através do seguinte resultado:

Resultado 2.2 - O estimador de máxima verossimilhança da especifici-

dade, quando os parâmetros S_E e π são fixos, é dado por

$$\widehat{E_S} = \begin{cases} 0, & \text{se } \dfrac{\left(\dfrac{n}{m+n}\right)+(S_E-1)\pi}{(1-\pi)} < 0, \\ \dfrac{\left(\dfrac{n}{m+n}\right)+(S_E-1)\pi}{(1-\pi)}, & \text{se } 0 \leq \dfrac{\left(\dfrac{n}{m+n}\right)+(S_E-1)\pi}{(1-\pi)} \leq 1, \\ 1, & \text{se } \dfrac{\left(\dfrac{n}{m+n}\right)+(S_E-1)\pi}{(1-\pi)} > 1. \end{cases}$$

Demonstração: Aplicando o logaritmo na função de verossimilhança 2.1, temos

$$\begin{aligned} l(\pi, S_E, E_S|m+n, m, n) &= \ln\left[\ (S_E\pi + (1-E_S)(1-\pi))^m\ ((1-S_E)\pi + E_S(1-\pi))^n\ \right] \\ &= m\ln(S_E\pi + (1-E_S)(1-\pi)) + n\ln((1-S_E)\pi + E_S(1-\pi)). \end{aligned}$$

Supondo fixos os parâmetros S_E e π, derivamos a função acima em relação a E_S, igualamos a mesma a zero e encontramos o EMV de E_S, ou seja,

$$\begin{aligned} &\frac{\partial l(\pi, S_E, E_S|m+n, m, n)}{\partial E_S} = 0 \\ &\frac{m(-1+\pi)}{S_E\pi + (1-E_S)(1-\pi)} + \frac{n(1-\pi)}{(1-S_E)\pi + E_S(1-\pi)} = 0 \\ &\Rightarrow \frac{m(-1+\pi)}{S_E\pi + (1-E_S)(1-\pi)} = \frac{-n(1-\pi)}{(1-S_E)\pi + E_S(1-\pi)} \\ &\Rightarrow \frac{m}{S_E\pi + (1-E_S)(1-\pi)} = \frac{n}{(1-S_E)\pi + E_S(1-\pi)} \end{aligned}$$

$$\Rightarrow m(\pi - S_E\pi + E_S - E_S\pi) = n(S_E\pi + 1 - \pi - E_S + E_S\pi)$$

$$\Rightarrow m\pi - m\pi S_E + mE_S - m\pi E_S = n\pi S_E + n - n\pi - nE_S + n\pi E_S$$

$$\Rightarrow mE_S - m\pi E_S + nE_S - n\pi E_S = n\pi S_E + n - n\pi - m\pi + m\pi S_E$$

$$\Rightarrow E_S(m - m\pi + n - n\pi) = \pi S_E(m+n) - \pi(m+n) + n$$

$$\Rightarrow E_S((m+n)-\pi(m+n)) = (S_E\pi-\pi)(m+n)+n$$

$$\Rightarrow E_S((m+n)(1-\pi)) = (S_E\pi-\pi)(m+n)+\frac{n(m+n)}{(m+n)}$$

$$\Rightarrow E_S(1-\pi) = \pi(S_E-1)+\frac{n}{(m+n)}$$

$$\Rightarrow \widehat{E_S} = \frac{(S_E-1)\pi+\left(\frac{n}{m+n}\right)}{(1-\pi)}.$$

Logo,

$$\widehat{E_S} = \frac{\left(\frac{n}{m+n}\right)+(S_E-1)\pi}{(1-\pi)}. \qquad \blacksquare$$

2.2.3 Estimador de máxima verossimilhança da π

Segundo Saraiva (2004), o EMV para o parâmetro π é obtido através do seguinte resultado:

Resultado 2.3 - O estimador de máxima verossimilhança da prevalência, quando os parâmetros S_E e E_S são fixos, é dado por

$$\widehat{\pi} = \begin{cases} 0, & \text{se } \dfrac{\left(\frac{m}{m+n}\right)-(1-E_S)}{(S_E+E_S-1)} < 0, \\ \dfrac{\left(\frac{m}{m+n}\right)-(1-E_S)}{(S_E+E_S-1)}, & \text{se } 0 \leq \dfrac{\left(\frac{m}{m+n}\right)-(1-E_S)}{(S_E+E_S-1)} \leq 1, \\ 1, & \text{se } \dfrac{\left(\frac{m}{m+n}\right)-(1-E_S)}{(S_E+E_S-1)} > 1. \end{cases}$$

Demonstração: Aplicando o logaritmo na função de verossimilhança 2.1, temos

$$\begin{aligned} l(\pi, S_E, E_S|m+n, m, n) &= \ln\left[\ (S_E\pi+(1-E_S)(1-\pi))^m\ ((1-S_E)\pi+E_S(1-\pi))^n\ \right] \\ &= m\ln(S_E\pi+(1-E_S)(1-\pi))+n\ln((1-S_E)\pi+E_S(1-\pi)). \end{aligned}$$

Supondo fixos os parâmetros S_E e E_S, derivamos a função acima em relação a π, igualamos a mesma a zero e encontramos o EMV de π, ou seja,

$$\frac{\partial l(\pi, S_E, E_S|m+n, m, n)}{\partial \pi} = 0$$

$$\frac{m(S_E - 1 + E_S)}{S_E\pi + (1-E_S)(1-\pi)} + \frac{n(1-S_E-E_S)}{(1-S_E)\pi + E_S(1-\pi)} = 0$$

$$\Rightarrow m(S_E-1+E_S)(\pi-S_E\pi+E_S-E_S\pi) = -n(1-S_E-E_S)(S_E\pi+1-\pi-E_S+E_S\pi)$$

$$\Rightarrow m(\pi - S_E\pi + E_S - E_S\pi) = n(S_E\pi + 1 - \pi - E_S + E_S\pi)$$

$$\Rightarrow m(\pi(1-S_E-E_S) + E_S) = n(\pi(S_E - 1 + E_S) + 1 - E_S)$$

$$\Rightarrow m\pi(1-S_E-E_S) + mE_S = n\pi(S_E - 1 + E_S) + n - nE_S$$

$$\Rightarrow m\pi(1-S_E-E_S) + mE_S = -n\pi(1 - S_E - E_S) + n - nE_S$$

$$\Rightarrow m\pi(1-S_E-E_S) + n\pi(1-S_E-E_S) = n - nE_S - mE_S$$

$$\Rightarrow \pi(1-S_E-E_S)(m+n) = n - nE_S - mE_S$$

$$\Rightarrow \widehat{\pi} = \frac{n - nE_S - mE_S}{(m+n)(1-S_E-E_S)} = \frac{-n + nE_S + mE_S}{(m+n)(1-S_E-E_S)(-1)}$$

$$= \frac{-n + nE_S + mE_S}{(m+n)(S_E + E_S - 1)} = \frac{m - m - n + nE_S + mE_S}{(m+n)(S_E + E_S - 1)}$$

$$= \frac{\dfrac{m - (m+n)(1-E_S)}{m+n}}{(S_E + E_S - 1)} = \frac{\left(\dfrac{m}{m+n}\right) - (1-E_S)}{(S_E + E_S - 1)}.$$

Logo,

$$\widehat{\pi} = \frac{\left(\dfrac{m}{m+n}\right) - (1-E_S)}{S_E + E_S - 1}. \qquad \blacksquare$$

Capítulo 3

A Curva *ROC* para Testes Diagnósticos

Existem vários índices de desempenho que podem ser utilizados na avaliação de sistemas de auxílio ao diagnóstico. Para tal, uma das medidas possíveis, e muito utilizada, é a análise da curva característica de resposta do observador (*ROC*), introduzida no contexto de imagens médicas por Metz (1986), que é definida como um procedimento estatístico que leva em conta o aspecto subjetivo envolvido em um determinado evento (Evans, 1981).

3.1 Histórico

Segundo Metz (1986), a análise *ROC* teve origem na teoria da decisão estatística, desenvolvida para avaliar a detecção de sinais em radar e na psicologia sensorial.

Durante a Segunda Guerra Mundial, a teoria da detecção do sinal começou a ser utilizada pelos operadores de radares. Neste contexto surgiu a curva *ROC*, que é uma ferramenta destinada a descrever o desempenho desses operadores (chamados de *receiver operators*). A curva *ROC* tinha como objetivo quantificar a habilidade dos operadores em distinguir um sinal de um ruído (Reiser e Faraggi, 1997). Essa habilidade ficou conhecida como *receiver operating characteristic*

(ROC). Quando o radar detectava algo se aproximando, o operador decidia se o que foi captado era um avião inimigo (sinal), ou algum outro objeto irrelevante, como por exemplo, uma nuvem ou um bando de aves (ruído) (Collinson, 1998).

Durante a detecção do sinal, o problema envolvendo sinais eletromagnéticos na presença de ruídos foi tratado como um problema de teste de hipóteses. A hipótese nula (H_0) foi identificada como sendo o ruído, enquanto a hipótese alternativa (H_1) estava associada ao ruído mais o sinal. O *erro tipo I* era um alarme qualquer falso, enquanto o *erro tipo II* eram as falhas, ambos perigosos em relação à defesa.

Na teoria de detecção do sinal, cabe ao observador decidir, com base na aleatoriedade, qual dos estímulos é resultado somente do ruído ou do ruído mais o sinal. Assim, o problema da detecção do sinal pode ser visto da seguinte forma:

- Existe uma ocorrência aleatória de dois acontecimentos: ruído e sinal;
- Cada acontecimento ocorre num intervalo de tempo bem definido;
- A evidência relativa a cada acontecimento (ou estímulo físico) varia de experiência para experiência, e tem um resultado que vem a ser a representação probabilística do acontecimento;
- Após cada observação, o operador deve tomar uma decisão do tipo: sim, o sinal encontra-se presente, ou não, o sinal encontra-se ausente, ou seja, tem-se evidências somente sobre o ruído.

3.2 Curva *ROC*

A teoria da análise *ROC* teve como base a idéia de que para qualquer sinal sempre existirá um fundo ruidoso que varia aleatoriamente sobre um valor médio.

Segundo Evans (1981), quando um estímulo está presente, a atividade que este cria no sistema de obtenção de imagem é adicionada ao ruído existente

naquele momento. Este ruído pode estar dentro do próprio sistema ou fazer parte do padrão de entrada.

A tarefa do observador (ou sistema automático) é determinar se o nível de atividade no sistema é devido apenas ao ruído ou ao resultado de um estímulo adicionado ao ruído. A tarefa de diagnosticar consiste na apresentação de imagens contendo ou não uma anormalidade associada a um observador que deve responder se existe ou não uma anormalidade presente.

O desempenho de um teste descrito em auxiliar diagnósticos clínicos é usualmente demonstrado pela curva *ROC*. A maior vantagem da curva *ROC* está em sua simplicidade, pois esta é uma representação visual direta do desempenho de um teste, de acordo com o conjunto de suas possíveis respostas.

Na curva *ROC* não há necessidade de se obter amostras que refletem a prevalência da doença, pois, em geral, a prevalência de muitas doenças é expressa por uma proporção relativamente pequena de indivíduos doentes (5% do total de indivíduos envolvidos no estudo já é suficiente).

Podemos definir curva *ROC* de duas formas diferentes, sendo uma delas mais restrita, em termos da razão de verossimilhança, e outra mais geral, em termos da variável de decisão.

Definição em termos da variável de decisão G

Uma curva *ROC* é o conjunto possível de matrizes 2×2, que resulta quando intervalos disjuntos do eixo das abscissas são sucessivamente adicionados ao intervalo de aceitação, onde a inclusão de intervalos começa com o intervalo vazio e termina com todo o eixo das abscissas. Os conjuntos possíveis de matrizes 2×2 estão restringidos pelas duas distribuições de G (Egan, 1975).

Definição em termos da função de verossimilhança $l(G)$

Uma curva *ROC* é o conjunto possível de matrizes 2×2, que resulta quando um valor de corte $r = l(G_o)$ varia de uma forma contínua de seu maior valor possível até o menor valor possível. Este conjunto de matrizes 2×2 é único

para as duas distribuições de G (Egan, 1975).

3.2.1 Regra de decisão - Ponto de corte

Segundo Martinez (2001), muitos testes diagnósticos não produzem resultados que são expressos como os mostrados na Figura 2.1, mas produzem uma resposta sob forma de uma variável categorizada ordinal ou contínua. Em casos como este, é empregada uma regra de decisão baseada em buscar um ponto de corte que resume tal quantidade em uma resposta dicotômica, de forma que um indivíduo com mensurações maiores que o ponto de corte é classificado como doente. Analogamente, um indivíduo com mensurações menores que o ponto de corte é classificado como não doente. Assim, para diferentes pontos de corte, dentro da amplitude dos possíveis valores que o teste que está sendo investigado produz, podemos estimar S_E e E_S. Um gráfico dos resultantes pares de S_E e $1 - E_S$ constitui uma curva *ROC* (Altman e Bland, 1994).

Seja R uma variável aleatória que representa o resultado de um teste diagnóstico, em que devemos considerar a seguinte regra de decisão baseada em um valor r_0 (ponto de corte): se $R > r_0$, o indivíduo é classificado como positivo (doente), e se $R \leq r_0$, o indivíduo é classificado como negativo (não doente).

Para um ponto de corte r_0 qualquer, temos que $S_E = P(Y > r_0)$ e $E_S = P(X \leq r_0)$, onde Y e X são variáveis aleatórias que representam os possíveis resultados do teste diagnóstico para indivíduos doentes e não doentes, respectivamente. Assim, a curva *ROC* é uma função contínua de S_E *versus* $1-E_S$, para diversos valores de r_0 obtidos dentro do espaço amostral de R, ligados por retas. Então, cada ponto de corte é associado a um único par $(1 - E_S\ ,\ S_E)$.

Observação: Para qualquer r_0 pertencente ao espaço amostral R, S_E seria igual a E_S, isto é, o teste sob investigação teria a mesma proporção de resultados positivos tanto para doentes quanto para não doentes. Logo, usamos $1 - E_S$ e não somente E_S no eixo das abscissas.

De uma maneira geral, quanto menor for o ponto de corte, maior será a habilidade do teste em classificar os doentes como positivos, isto é, maior será S_E.

Em determinadas situações, quando não pode correr o risco de não diagnosticar determinada doença, é melhor privilegiar S_E. Assim, os testes sensíveis são utilizados para rastrear a doença em grupos populacionais. Exemplos: durante o rastreamento do câncer, é indesejável que alguns indivíduos doentes deixem de ser detectados, sendo assim tratados em estágios menos avançados da doença; o uso do teste anti-HIV em bancos de sangue.

Por outro lado, é inevitável que indivíduos não doentes sejam classificados como positivos, isto é, menor será E_S. Em situações em que, por exemplo, o diagnóstico (resultado falso positivo) possa trazer prejuízos ao indivíduo (emocional, físico ou financeiro), é melhor privilegiar E_S (falso diagnóstico). Assim, os testes específicos são utilizados para confirmar um diagnóstico, pois um teste bastante específico raramente resultará positivo na ausência da doença. Exemplo: o teste anti-HIV.

Segundo Fletcher *et al.* (1996), um teste muito sensível raramente deixará de diagnosticar indivíduos com a doença e um teste muito específico raramente classificará como doente um indivíduo sem a doença.

O número de pontos de corte escolhidos para construção da curva ROC varia de acordo com a necessidade do clínico. Contudo, sabemos que uma quantidade maior de pontos de corte escolhidos para estimar a curva ROC fornecem um resultado mais preciso. Assim sendo, existem três tipos de curva ROC baseadas nas quantidades de pontos de corte estabelecidos.

A curva ROC **empírica** é baseada em todos os possíveis pontos de corte que a amostra permite estabelecer, relativos a cada medida da amostra, sendo no máximo $m + n$ pontos de corte.

A curva ROC **aproximada** tem o número de pontos de corte estabelecidos pelo usuário, sendo menores que $m + n$.

A curva ROC **teórica** é aquela onde é conhecido o modelo estatístico que determina a distribuição da variável aleatória referente ao resultado do teste diagnóstico nos indivíduos doentes e não doentes.

Perante os critérios adotados, vamos entender o significado da localização

de um ponto de corte na curva *ROC*.

Podemos definir um critério estrito (por exemplo, apenas se designa o paciente positivo quando a evidência da doença é muito forte) como sendo aquele que conduz a uma pequena $1 - E_S$ e também a uma, relativamente, pequena S_E, isto é, gera um ponto na curva *ROC* que se situa no canto inferior esquerdo. Progressivamente, critérios menos estritos conduzem a maiores frações de ambos os tipos, ou seja, pontos situados no canto superior direito da curva no espaço *ROC*. Essa situação pode ser descrita graficamente pela curva *ROC* apresentada na Figura 3.1.

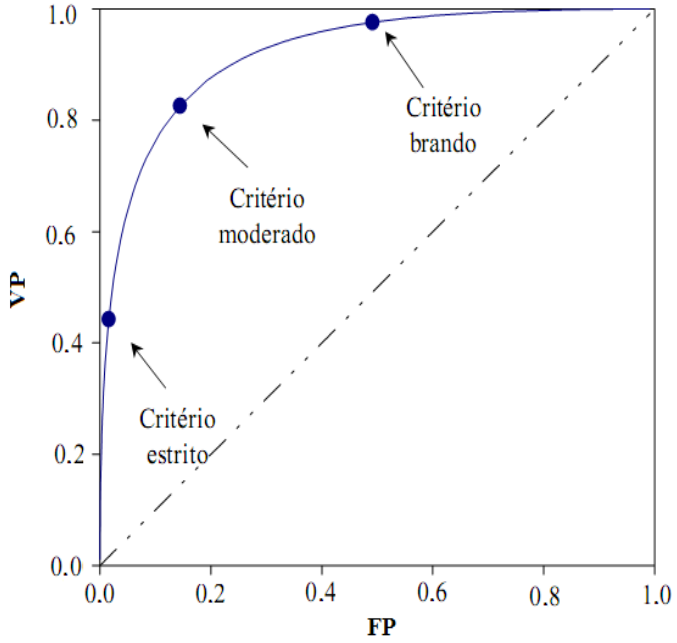

FIGURA 3.1: Curva *ROC* para uma dada capacidade de discriminação conforme a variação do critério de decisão.

O valor do ponto de corte é o que define a região de rejeição para a curva *ROC* que ilustra um teste diagnóstico, isto é, define as dimensões do *Erro tipo I* e do *Erro tipo II*.

Aplicamos, para testes diagnósticos, as hipóteses apresentadas no início deste Capítulo sobre a teoria da detecção do sinal.

Seja um teste de hipóteses dado por

$$\begin{cases} H_0: & \text{O indivíduo é doente } (D = 1), \\ H_1: & \text{O indivíduo não é doente } (D = 0). \end{cases}$$

Temos que

$$Prob\ (Erro\ tipo\ I) = P(Rejeitar\ H_0 \mid H_0\ \ Verdadeira) = P(R = 0 \mid D = 1)$$
$$= 1 - P(R = 1 \mid D = 1) = 1 - S_E,$$

e

$$Prob\ (Erro\ tipo\ II) = P(Não\ rejeitar\ H_0 \mid H_0\ Falsa) = P(R = 1 \mid D = 0)$$
$$= 1 - P(R = 0 \mid D = 0) = 1 - E_S.$$

Logo, conforme o valor do ponto de corte varia, os erros também vão variando de valor (à medida que um aumenta, o outro diminui de valor, e vice-versa).

Sob o ponto de vista da teoria de teste de hipóteses, uma curva *ROC* é conceitualmente equivalente a uma curva que mostra a relação entre a potência do teste e a probabilidade de cometer um *erro tipo I* com a variação do valor crítico do teste estatístico.

A curva *ROC* descrita até agora é representada no plano unitário. No entanto, existe uma outra forma para a representação desta curva, que é no plano binormal (ver Apêndice A).

É bom ressaltar que, para as inferências estatísticas, como por exemplo, comparações de várias curvas *ROC* num mesmo gráfico, são frequentemente usadas curvas *ROC* empíricas, pois estas estimam com maior precisão a curva *ROC* teórica, que usualmente é desconhecida.

A curva *ROC* não sofre mudanças se as observações amostrais forem submetidas a transformações monótonas, como por exemplo, logaritmo e raiz quadrada.

3.2.2 Medida de precisão - Área sob a curva *ROC*

Um teste diagnóstico que gera a mesma proporção de resultados positivos para indivíduos doentes e para indivíduos não doentes ($S_E = 1 - E_S$), na prática, não tem utilidade. Já um teste totalmente capaz de discriminar indivíduos doentes dos não doentes teria, para algum ponto de corte, uma sensibilidade de 100% e especificidade de 100%, fato este representado no canto superior esquerdo da curva. Em outras palavras, o valor da S_E sendo 1, significa acertar 100% dos indivíduos doentes; analogamente, o valor da E_S sendo 1, significa acertar 100% dos indivíduos não doentes (Martinez, 2001).

O fato de podermos afirmar que o teste é ou não capaz de discriminar indivíduos doentes dos não doentes esta diretamente ligado a uma medida de precisão da curva *ROC*, que é a área sob essa curva.

A área sob a curva *ROC* é uma medida que resume o desempenho de um teste diagnóstico, pois estimamos a mesma considerando todas as E_S e S_E relativas a cada ponto de corte r_0 estipulado.

A precisão de um teste diagnóstico depende de quão bem ocorre a separação entre doentes e não doentes. A precisão é medida pela área sob a curva *ROC*. Uma área igual a 1 representa um teste perfeito, ou seja, o teste tem que acertar todos os diagnósticos da doença. Uma área igual a 0.5 representa um teste sem valor, pois o simples lançamento de uma moeda levaria a uma curva próxima da reta identidade, fato este que proporciona área 0.5. Logo, quanto maior a área sob a curva *ROC*, melhor é a classificação do teste conforme seu desempenho.

Computar a área sob a curva *ROC* necessita do conhecimento da função $ROC(r_0)$. Para isso, métodos computacionais são utilizados, tais como:

- Método paramétrico usando um estimador de máxima verossimilhança para aproximar uma curva suave para os valores dos pares ordenados $(1-E_S\ ,\ S_E)$;

- Método não paramétrico baseado na construção de trapézios sob a curva.

Regra do Trapézio

Segundo Bamber (1975), a área sob a curva *ROC* empírica, traçada nos $T = 1, ..., t$ $(n + m = t)$ valores para o ponto de corte r_t, pode ser estimada pela soma das áreas dos $t - 1$ trapézios que dividem a curva.

Seja R uma variável aleatória que representa o resultado de um teste diagnóstico. Sejam Y e X as variáveis aleatórias que representam os valores de R para indivíduos doentes e não doentes, respectivamente. Sem perda de generalidade, vamos assumir que Y e X são variáveis discretas.

A área do t-ésimo trapézio é dada por $A_t = a_{t1} + a_{t2}$, onde a_{t1} é um triângulo cuja área é $\frac{1}{2}\,[P(X \leq r_t) - P(X \leq r_{t-1})]\,[P(Y \leq r_t) - P(Y \leq r_{t-1})]$ e a_{t2} é um retângulo cuja área é $[P(X \leq r_t) - P(X \leq r_{t-1})]\,[P(Y \leq r_{t-1})]$, ou seja,

$$\begin{aligned} A_t &= [P(X \leq r_t) - P(X \leq r_{t-1})]\,[P(Y \leq r_{t-1})] + \\ &\quad \frac{1}{2}\,[P(X \leq r_t) - P(X \leq r_{t-1})]\,[P(Y \leq r_t) - P(Y \leq r_{t-1})] \\ &= P(X = r_t)P(Y \leq r_{t-1}) + \frac{1}{2}\,P(X = r_t)P(Y = r_t) \\ &= P(X = r_t)[P(Y \leq r_{t-1}) + \frac{1}{2}\,P(Y = r_t)], \end{aligned}$$

como apresentada na Figura 3.2.

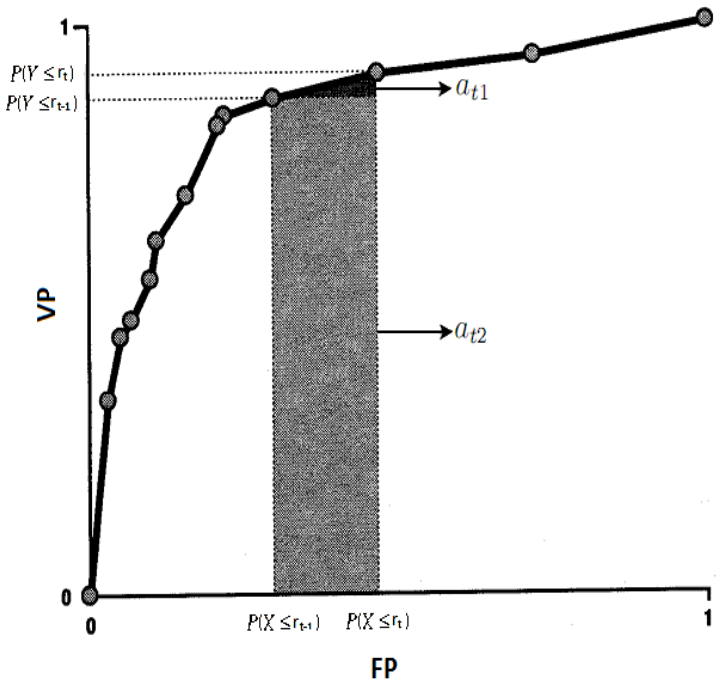

FIGURA 3.2: Área do t-ésimo trapézio que é dada por $A_t = a_{t1} + a_{t2}$.

Assim, a área total sob a curva ROC empírica pode ser estimada pela soma das áreas de todos os trapézios, no qual

$$\begin{aligned} A(X,Y) &= \sum_{t=1}^{t}\left(P(X=r_t)[P(Y \leq r_{t-1}) + \frac{1}{2}\,P(Y=r_t)]\right) \\ &= \sum_{t=1}^{t} P(X=r_t)P(Y \leq r_{t-1}) + \frac{1}{2}\sum_{t=1}^{t} P(X=r_t)\;P(Y=r_t) \\ &= P(Y<X) + \frac{1}{2}\,P(Y=X). \end{aligned}$$

Sendo X e Y variáveis contínuas, podemos generalizar esse resultado, pois estimamos S_E e E_S para cada uma das observações amostrais, sendo assim tratadas tal como observações de uma variável discreta. Logo, a área sob a curva ROC é dada por

$$\begin{aligned} A(X,Y) &= \int_0^1 P(Y \leq t)dP(X \leq r) \\ &= \int_{-\infty}^{+\infty} P(Y \leq r) f_X(r)dr \\ &= P(Y \leq X), \end{aligned}$$

em que $f_X(r)$ é a função densidade de X.

Quando X e Y são variáveis contínuas, então $P(X = Y) = 0$.

3.3 Aplicação

Vamos apresentar um exemplo de teste diagnóstico, em que a curva *ROC* destinada a representar a eficiência do mesmo está associada a uma variável aleatória categórica ordinal (discreta).

Um estudo tem como objetivo avaliar o desempenho de critérios morfológicos e *colordopplervelocimétricos* para auxiliar o diagnóstico de tumores mamários malígnos (Marussi, 2001 e Martinez *et al.*, 2003). Para este, foram selecionadas 388 mulheres portadoras de nódulos sólidos na mama, pondendo ser palpáveis ou não. Estes nódulos foram identificados por radiografia ou ultrasonografia no Serviço de Ultra-Sonografia do Centro de Atenção Integral à Saúde da Mulher da Universidade Estadual de Campinas. Nesta aplicação, excluímos as pacientes portadoras de dois ou mais nódulos. As mulheres selecionadas formam uma amostra de $n = 246$ portadoras de tumores benignos e $m = 142$ portadoras de tumores malignos.

No decorrer desta aplicação, entendemos como indivíduos não doentes aqueles portadores de tumores benignos e indivíduos doentes os portadores de tumores malignos.

O padrão-ouro foi baseado no diagnóstico histológico do nódulo. Um dos critérios *colordopplervelocimétricos* considerados foi o índice de cor, avaliado subjetivamente da seguinte forma: ausente, até $1/4$, até $1/2$ e maior que a metade da área do nódulo. Esses valores são dados em relação à área total do nódulo que é ocupada por vasos, representados por pontos coloridos identificados na radiografia.

A Tabela 3.1 resume as informações encontradas a partir da forma subjetiva de classificação.

TABELA 3.1: Distribuição dos pacientes portadores de tumores malignos e benignos

Área do nódulo ocupada	**Tumores malignos (m)**	**Tumores benignos (n)**
Ausente	19	120
Até 1/4	44	85
Até 1/2	58	34
Maior que 1/2	21	7

Com os dados apresentados conforme a regra de classificação, para este específico teste diagnóstico, encontramos cinco pontos de corte fixos, sendo dois deles os extremos da curva, que são os pontos (0 , 0) e (1 , 1). Logo, três pontos de corte "intermediários" servirão, juntamente com os dois pontos extremos, para traçar a curva *ROC*, sendo estes denominados por **A**, **B** e **C**.

Para o ponto de corte **A**, segundo o índice de cor, consideramos a malignidade positiva (+) para resultados onde o aparelho indica a presença de vasos na extensão do nódulo e negativa (−) a ausência desses. Os resultados encontrados estão dispostos na Tabela 3.2.

TABELA 3.2: Ponto de corte **A**

	Maligno	**Benigno**
Presente (+)	123	126
Ausente (-)	19	120
	$S_E = 0.8661972$	$E_S = 0.4878049$

Estimamos S_E e E_S para o ponto de corte **A** da seguinte forma

$$\widehat{S_E} = \frac{VP}{VP + FN} = \frac{123}{123 + 19} = \frac{123}{142} = 0.8661972,$$

e

$$\widehat{E_S} = \frac{VN}{FP + VN} = \frac{120}{126 + 120} = \frac{120}{246} = 0.4878049.$$

Entretanto, no momento em que construímos a curva *ROC* a partir dos pontos de corte estabelecidos, não podemos esquecer do fato de que essa curva tem como abscissa os valores de $1 - E_S$ e não simplesmente os valores de E_S.

Assim, devemos calcular $1 - E_S = 1 - 0.4878049 = 0.5121951$ e, a partir deste novo valor encontrado, estabelecer as coordenadas cartesianas para o ponto de corte **A**, como sendo $\mathbf{A} = (0.5121951\ ,\ 0.8661972)$.

Para o ponto de corte **B**, segundo o índice de cor, consideramos a malignidade positiva (+) para resultados onde o aparelho indica a presença de vasos em maior que 1/4 da área do nódulo e negativa (−) para a ausência ou até 1/4 da área. Os resultados encontrados estão dispostos na Tabela 3.3.

TABELA 3.3: Ponto de corte **B**

	Maligno	**Benigno**
> que 1/4 **(+)**	79	41
Ausente ou até 1/4 **(-)**	63	205
	$S_E = 0.556338$	$E_S = 0.8333333$

Estimamos S_E e E_S para o ponto de corte **B** da seguinte forma

$$\widehat{S_E} = \frac{VP}{VP + FN} = \frac{79}{142} = 0.556338,$$

e

$$\widehat{E_S} = \frac{VN}{FP + VN} = \frac{205}{246} = 0.8333333.$$

Assim, a coordenada cartesiana para este ponto de corte é $\mathbf{B} = (0.1666667\ ,\ 0.556338)$.

Para o ponto de corte **C**, segundo o índice de cor, consideramos a malignidade positiva (+) para resultados onde o aparelho indica a presença de vasos em maior que 1/2 da área do nódulo e negativa (−) para a ausência ou até 1/2 da área. Os resultados encontrados estão dispostos na Tabela 3.4.

TABELA 3.4: Ponto de corte **C**

	Maligno	**Benigno**
> que 1/2 **(+)**	21	7
Ausente ou até 1/2 **(-)**	121	239
	$S_E = 0.1478873$	$E_S = 0.9715447$

Estimamos S_E e E_S para o ponto de corte **C** da seguinte forma

$$\widehat{S_E} = \frac{VP}{VP + FN} = \frac{21}{142} = 0.1478873,$$

e

$$\widehat{E_S} = \frac{VN}{FP + VN} = \frac{239}{246} = 0.9715447.$$

Assim, a coordenada cartesiana para este ponto de corte é $\mathbf{C} = (0.02845528\ ,\ 0.1478873)$.

Logo, a curva *ROC* resultante é o gráfico que une com retas os pares $(1 - E_S\ ,\ S_E)$ estimados para cada ponto de corte.

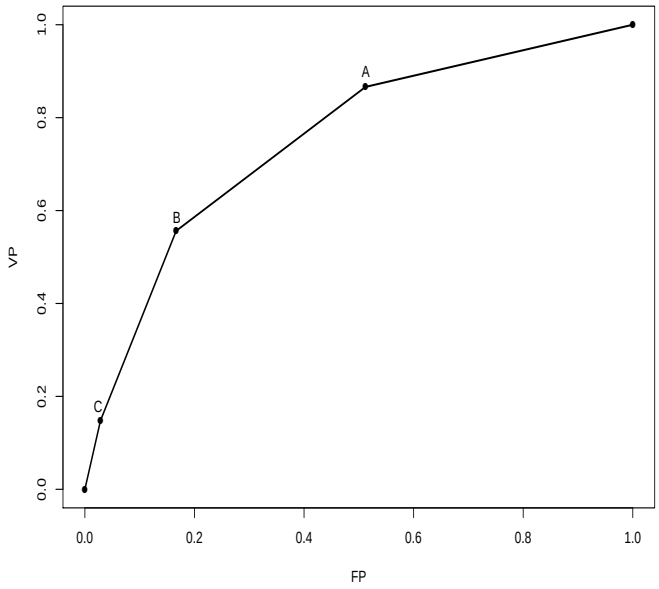

FIGURA 3.3: Curva *ROC* para o teste diagnóstico.

Vamos calcular a área sob curva *ROC* encontrada. Esta será estimada pelo método dos trapézios, onde dividimos a curva conforme seus respectivos

pontos de corte. Não temos que encontrar as probabilidades descritas na teoria sobre área sob a curva *ROC*, pois tratamos de uma curva *ROC* aproximada, e não uma curva *ROC* empírica.

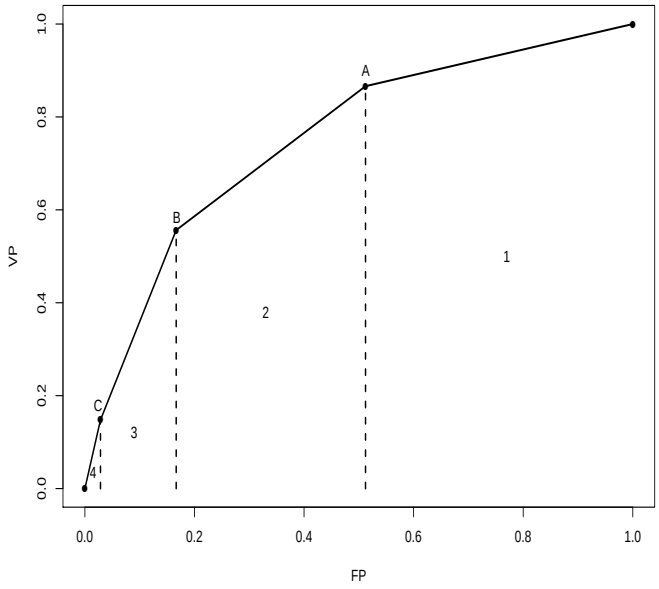

FIGURA 3.4: Divisão da curva *ROC* para o cálculo da área.

Na Figura 3.4, levando em consideração que 1, 2 e 3 são trapézios e que 4 é um triângulo, temos que a área total é dada por

$$\begin{aligned} A &= A_1 + A_2 + A_3 + A_4 \\ &= \frac{(1 + 0.8661972)(1 - 0.5121951)}{2} + \frac{(0.8661972 + 0.556338)(0.5121951 - 0.1666667)}{2} \\ &+ \frac{(0.556338 + 0.1478873)(0.1666667 - 0.02845528)}{2} + \frac{(0.1478873)(0.02845528)}{2} \\ &= 0.4551700 + 0.2457632 + 0.04866598 + 0.002104087 \\ &= 0.7517033 \cong 75.17\%. \end{aligned}$$

Para interpretarmos a área sob a curva *ROC* deste teste diagnóstico (designado de agora em diante por I), vamos considerar um outro teste diagnóstico (designado por II), do mesmo padrão que o primeiro. Utilizando o *Software R 2.8.1*, fixamos uma semente no valor de 830127, e aplicamos a técnica de reamostragem *bootstrap* nos dados do teste I. A Tabela 3.5 resume as informações

encontradas para o teste diagnóstico II.

TABELA 3.5: Distribuição dos pacientes portadores de tumores malignos e benignos para o teste diagnóstico II

Área do nódulo ocupada	**Tumores malignos (m)**	**Tumores benignos (n)**
Ausente	19	120
Até 1/4	50	88
Até 1/2	60	36
Maior que 1/2	13	2

Logo, com os dados apresentados na Tabela 3.5, encontramos os seguintes pontos de corte: **D** = (0.5121951 , 0.8661972), **E** = (0.1544715 , 0.5140845) e **F** = (0.008130081 , 0.0915493). A curva *ROC* para este teste é apresentada na Figura 3.5.

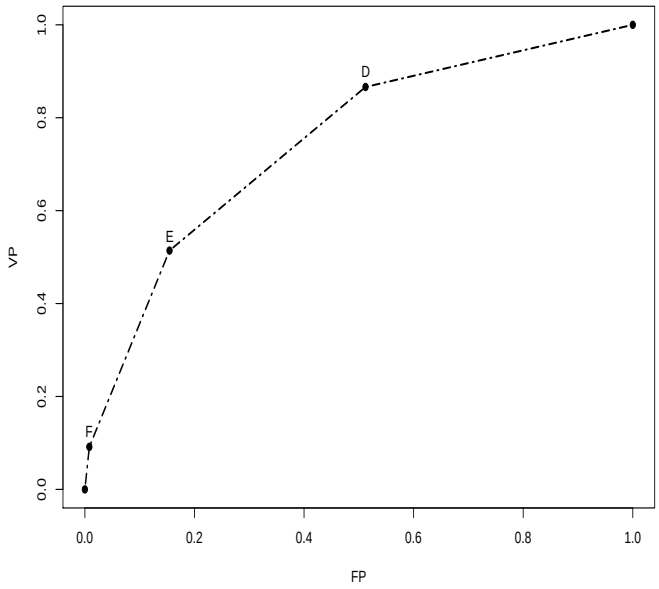

FIGURA 3.5: Curva *ROC* para o teste diagnóstico II.

Fazendo os cálculos, obtemos o valor de 0.7480363 (74.80%) para a área sob a curva *ROC* que representa o teste diagnóstico II.

Na Figura 3.6, temos as curvas *ROC* para os dois testes diagnósticos.

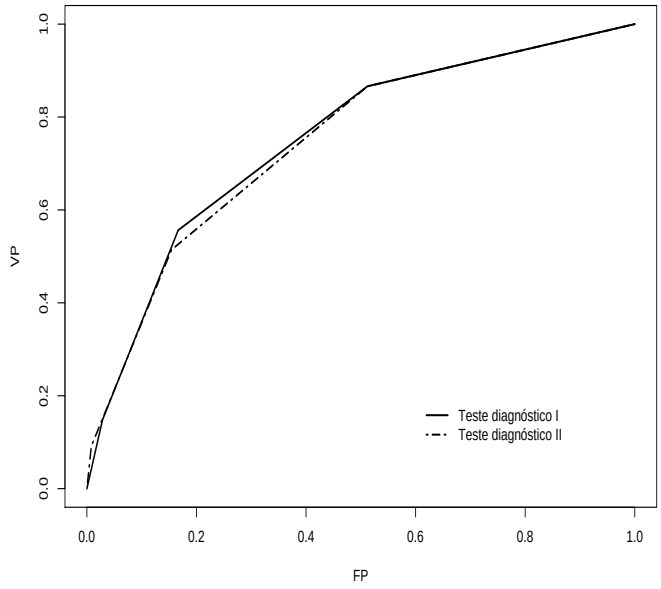

FIGURA 3.6: Curvas *ROC* para os testes diagnósticos I e II.

Quando comparadas as duas áreas, ou seja, a área sob a curva *ROC* que representa o teste diagnóstico I (A_I) com a área sob a curva *ROC* que representa o teste diagnóstico II (A_{II}), notamos que $A_I = 0.7517 > 0.7480 = A_{II}$. Assim, para este específico conjunto de dados e para esta específica reamostragem, o teste diagnóstico I apresenta um desempenho um pouco mais preciso em relação ao teste II, pois sua área é ligeiramente maior. Logo, o teste I consegue fazer uma melhor distinção entre tumores malignos e benignos, diminuindo a chance do clínico em dar um diagnóstico errado para essa determinada enfermidade. Porém, não podemos desconsiderar o teste diagnóstico II, que apresenta uma área de 0.7480, o que também leva a conclusão de um desempenho satisfatório.

Um outro fato que deve ser lembrado é que a mensuração da área do nódulo ocupada por vasos é baseada na visualização de pontos coloridos em uma imagem produzida por um aparelho, sendo assim bastante subjetiva.

Através deste exemplo, entendemos que a área é uma medida sob a curva *ROC* que mede a habilidade do teste em classificar corretamente indivíduos doentes e indivíduos não doentes.

Capítulo 4

Método *Bootstrap*

Elaborado por Efron (1979), o *bootstrap* é um processo de simulação via reamostragem utilizado na obtenção de estimativas pontuais e intervalares, bem como na avaliação da acurácia de estimativas e testes. Este método consiste, basicamente, na replicação do processo de estimação via reamostragem pela amostra ou distribuição da variável, caso seja conhecida, com parâmetros estimados via amostra.

Técnicas de reamostragem são úteis, em especial, quando for complicada a obtenção de estimadores por métodos analíticos. Devido à sua generalidade, a técnica *bootstrap* é adequada para soluções de problemas complexos.

Esta técnica foi desenvolvida inicialmente para fornecer a variabilidade das estimativas e medidas de vício. Porém, seu uso foi estendido para construção de intervalos de confiança, testes de hipóteses, entre outras análises.

Muitas vezes a distribuição de probabilidade é desconhecida. Nesse caso, o *bootstrap* é muito útil, pois é uma técnica que não exige diferentes fórmulas para cada problema e pode ser utilizada em casos gerais, não dependendo da distribuição original do parâmetro estudado.

Quando a distribuição do parâmetro a ser estimado é conhecida, a coincidência entre o intervalo paramétrico baseado na distribuição de probabilidade do parâmetro e o intervalo *bootstrap* reforçam a hipótese de veracidade a respeito das suposições do modelo paramétrico.

4.1 *Bootstrap*

A partir de uma amostra aleatória de tamanho v de uma população, é obtida, com reposição, uma reamostra de mesmo tamanho v dessa amostra original. Essa reamostra deve ser coletada de maneira planejada, uma vez que mal obtida não representa bem a população, fazendo com que a técnica não forneça resultados confiáveis.

Para que a aplicação da técnica *bootstrap* resulte em valores confiáveis, devem ser feitas, a partir da primeira reamostra, muitas reamostras do mesmo tamanho v. É importante que todas as reamostragens sejam realizadas selecionando os valores de forma aleatória e com reposição.

Cada amostra obtida por reamostragem é uma amostra *bootstrap*. O processo de reamostragem é executado inúmeras vezes, de modo que as estimativas dos parâmetros sejam obtidas. Essas estimativas geram uma distribuição denominada distribuição *bootstrap*.

Em alguns casos, a estimativa do parâmetro de interesse obtida na amostra original é utilizada como parâmetro da função de densidade da variável aleatória correspondente. Essa densidade é uma estimativa da verdadeira densidade populacional. Por meio dela, são realizadas amostragens e em cada etapa é obtida uma estimativa do parâmetro de interesse. Repetido esse processo um número grande de vezes, é gerada a distribuição.

A distribuição *bootstrap* tem aproximadamente a mesma forma e amplitude que a distribuição amostral, porém está centrada na estatística dos dados originais, enquanto a distribuição amostral está centrada no parâmetro da população.

Sejam $W_1, W_2, ..., W_v$ uma amostra aleatória simples, de tamanho v, com função distribuição de probabilidade F. Suponha que estamos interessados em inferir sobre uma quantidade populacional θ, onde $\theta = t(F)$, sendo que $\widehat{\theta}$ é o estimador para este parâmetro. Com o método de reamostragem *bootstrap*,

conseguimos inferir sobre a distribuição de $(\widehat{\theta} - \theta)$. Este fato acontece a partir da determinação de uma estimativa da distribuição F, sendo denominada, por exemplo, de $\tilde{F}$ e conhecida como função *plug-in*, de modo que $\widehat{\theta} = t(\tilde{F})$ (Taconeli e Barreto, 2005).

Segundo Canty (2000), usando reamostragem *bootstrap*, as réplicas de $\widehat{\theta}$, indicadas por $\widehat{\theta^*}$, tornam possível a aproximação em distribuição $(\widehat{\theta} - \theta) \approx (\widehat{\theta^*} - \widehat{\theta})$.

Este fato possibilita a estimação do viés (vício) e da variância de $\widehat{\theta}$, dados respectivamente por

$$\beta(\widehat{\theta}) = E(\widehat{\theta} - \theta \mid W \sim F) \approx E(\widehat{\theta}^* - \widehat{\theta} \mid W^* \sim \tilde{F}),$$

e

$$Var(\widehat{\theta}) = Var(\widehat{\theta} - \theta \mid W \sim F) \approx Var(\widehat{\theta}^* - \widehat{\theta} \mid W^* \sim \tilde{F}).$$

Sendo $\tilde{F}$ uma distribuição conhecida, torna-se fácil a determinação da distribuição $(\widehat{\theta}^* - \widehat{\theta} \mid W^* \sim \tilde{F})$. Esta aproximação pode acontecer através de métodos computacionalmente intensivos, como Monte Carlo, ou de maneira analítica.

O processo de reamostragem *bootstrap* pode ser não paramétrico ou paramétrico.

4.1.1 *Bootstrap* não paramétrico

O *bootstrap* não paramétrico consiste da utilização de $\widehat{F}$ como função *plug-in*. Isso é possível pela geração de reamostras com reposição e de mesmo tamanho da amostra original, a partir da distribuição empírica dos dados $\widehat{F}$. Essa função é dada por

$$\widehat{F}(w) = \frac{\#\{w_b \leq w\}}{v},$$

em que $\#\{\cdot\}$ é o número de vezes que a condição é verdadeira e $\widehat{F}(w)$ é a frequência relativa acumulada.

O *bootstrap* não paramétrico consiste na obtenção de amostras

$\mathbf{w}_b^* = (x_{b_1}^*, x_{b_2}^*, ..., x_{b_v}^*)$, onde $b = 1, 2, ..., B$, sendo w_b^* obtidas via reamostragem dos dados originais, responsáveis pela produção das B estimativas *bootstrap* $(\widehat{\theta}_1^*, \widehat{\theta}_2^*, ..., \widehat{\theta_B^*})$ (Taconeli, 2005).

4.1.2 *Bootstrap* paramétrico

O processo de reamostragem *bootstrap* paramétrico é baseado em modelos probabilísticos com parâmetros estimados via amostra original.

Seja F uma distribuição indexada por um parâmetro δ, com estimador $\widehat{\delta}$, baseado na amostra original $W_1, W_2, ..., W_v$, e considere $\widehat{\theta} = t(F_{\widehat{\delta}})$. Podemos considerar $F_{\widehat{\delta}}$ como estimador de F_δ.

Sejam $W_1^*, W_2^*, ..., W_B^*$ amostras da distribuição $F_{\widehat{\delta}}$ e $\widehat{\delta}_1^*, \widehat{\delta}_2^*, ..., \widehat{\delta_B^*}$ estimativas de $\widehat{\delta}$ para estas amostras.

O *bootstrap* paramétrico consiste, assim, na aproximação da distribuição $(t(F_{\widehat{\delta}}) - t(F_\delta))$ por $(t(F_{\widehat{\delta^*}}) - t(F_{\widehat{\delta}}))$ (Taconeli, 2005).

4.2 Intervalos de confiança *bootstrap*

Ao fazer inferência sobre o parâmetro θ, somente a estimativa pontual $\widehat{\theta}$ é imprecisa, pois não leva em consideração informações como a precisão do estimador ou o erro dessa estimação. A construção de intervalos de confiança leva em consideração esses valores, proporcionando assim a obtenção de estimativas mais confiáveis.

A acurácia e a precisão são as principais características para analisar o desempenho dos intervalos de confiança. A acurácia é o quão próximo a probabilidade de cobertura desses intervalos está do nível de confiança fixado e a precisão refere-se à variabilidade das estimativas encontradas. Para maiores informações sobre a acurácia, ver Efron e Tibshirani (1993).

A seguir, apresentaremos os intervalos de confiança *bootstrap* utilizados neste trabalho, durante a estimação das regiões de confiança de nosso interesse.

4.2.1 Intervalo *bootstrap* percentil

Efetuadas B reamostras *bootstrap* $(w_1^*, w_2^*, ..., w_B^*)$, as quais originaram B estimativas para o parâmetro de interesse $(\widehat{\theta}_1^*, \widehat{\theta}_2^*, ..., \widehat{\theta}_B^*)$, um intervalo de confiança com probabilidade de cobertura de $(1-\alpha)$ construído pelo método percentil é aquele dado pelos quantis $(\frac{\alpha}{2})$ e $(1-\frac{\alpha}{2})$ da função de distribuição acumulada de $\widehat{\theta}^*$, denominada de $\widehat{H}$. Este intervalo pode ser escrito da seguinte forma

$$\left[\widehat{H}^{-1}_{(\frac{\alpha}{2})} \ ; \ \widehat{H}^{-1}_{(1-\frac{\alpha}{2})}\right]. \tag{4.1}$$

Segundo Efron e Tibshirani (1993), como $\widehat{H}^{-1}_{(\frac{\alpha}{2})}$ é o $\frac{\alpha}{2}$-ésimo quantil da distribuição *bootstrap*, o intervalo apresentado na expressão 4.1 pode ser reescrito como

$$\left[\widehat{\theta}^*_{(\frac{\alpha}{2})} \ ; \ \widehat{\theta}^*_{(1-\frac{\alpha}{2})}\right]. \tag{4.2}$$

O intervalo apresentado na expressão 4.2 consiste na porção central de tamanho $(1-\alpha)$ da distribuição de $\widehat{\theta}^*$.

Esta é a situação ideal, pois está relacionada a um número infinito de replicações. Contudo, na prática, utilizamos um número finito B de reamostragens *bootstrap* e B estimativas *bootstrap*.

Logo, um intervalo de confiança do tipo *boostrap* percentil é dado por

$$\left[\widehat{\theta}^{*B}_{(\frac{\alpha}{2})} \ ; \ \widehat{\theta}^{*B}_{(1-\frac{\alpha}{2})}\right]. \tag{4.3}$$

4.2.2 Intervalo *bootstrap* normal

O intervalo *bootstrap* normal também é conhecido como intervalo *bootstrap* padrão.

Seja $\widehat{\theta}$ o estimador do parâmetro θ. Da aproximação de $(\widehat{\theta}-\theta)$ por uma normal, dada por

$$Z = \frac{\widehat{\theta}-\theta}{\sqrt{Var(\widehat{\theta})}} \sim N(0 \ , \ 1), \tag{4.4}$$

construímos um intervalo bilateral com $100(1-\alpha)\%$ de nível de confiança. Este intervalo é dado por

$$\left[\widehat{\theta} - z_{(1-\frac{\alpha}{2})}\sqrt{Var(\widehat{\theta})}\ ;\ \widehat{\theta} - z_{(\frac{\alpha}{2})}\sqrt{Var(\widehat{\theta})}\right], \tag{4.5}$$

sendo $z_{(1-\frac{\alpha}{2})}$ e $z_{(\frac{\alpha}{2})}$ os quantis da distribuição normal padrão.

O método *bootstrap* pode ser aplicado no cálculo do viés estimado ($\beta_B(\widehat{\theta})$) e no cálculo da variância do estimador ($Var_B(\widehat{\theta})$), produzindo assim um fator de correção para a estimativa pontual. Essas quantidades são obtidas pela geração de B amostras *bootstrap*, e são dadas por

$$\beta_B(\widehat{\theta}) = \frac{1}{B}\sum_{b=1}^{B}(\widehat{\theta}_b^* - \widehat{\theta}), \tag{4.6}$$

e

$$Var_B(\widehat{\theta}) = \frac{1}{B-1}\sum_{b=1}^{B}(\widehat{\theta}_b^* - \overline{\widehat{\theta}^*})^2, \tag{4.7}$$

em que $\widehat{\theta}_b^*$ corresponde a estimativa obtida através da b-ésima reamostra e $\overline{\widehat{\theta}}^*$ é a média dessas estimativas, dada por

$$\overline{\widehat{\theta}^*} = \frac{1}{B}\sum_{b=1}^{B}\widehat{\theta}_b^*.$$

Logo, um intervalo de confiança usando a aproximação normal, calculado pelo método de reamostragem *bootstrap*, é dado por

$$\left[\widehat{\theta} - \beta_B(\widehat{\theta}) - z_{(1-\frac{\alpha}{2})}\sqrt{Var_B(\widehat{\theta})}\ ;\ \widehat{\theta} - \beta_B(\widehat{\theta}) - z_{(\frac{\alpha}{2})}\sqrt{Var_B(\widehat{\theta})}\right]. \tag{4.8}$$

4.2.3 Intervalo *bootstrap* studentizado

A aproximação apresentada na expressão 4.4 tem bons resultados para amostras de tamanho grande.

Quando trabalhamos com amostras de tamanho reduzido, uma boa apro-

ximação é dada por

$$Z = \frac{\widehat{\theta} - \theta}{\sqrt{Var(\widehat{\theta})}} \sim t_{v-1}, \tag{4.9}$$

em que t_{v-1} se refere à distribuição t-Student com $(v-1)$ graus de liberdade.

Baseado na expressão 4.5 e na aproximação 4.9, um intervalo bilateral com $100(1-\alpha)\%$ de nível de confiança é dado por

$$\left[\widehat{\theta} - t_{(1-\frac{\alpha}{2}),v-1}\sqrt{Var(\widehat{\theta})}\ ;\ \widehat{\theta} - t_{(\frac{\alpha}{2}),v-1}\sqrt{Var(\widehat{\theta})}\right], \tag{4.10}$$

sendo $t_{(\alpha),v-1}$ o α-ésimo quantil da distribuição t-Student com $(v-1)$ graus de liberdade.

Para construirmos intervalos de confiança *bootstrap* a partir do intervalo apresentado na expressão 4.9, calculamos a estimativa $\widehat{\theta}_b^*$, que é proveniente da reamostra w_b^*, e a quantidade

$$Z_b^* = \frac{\widehat{\theta}_b^* - \widehat{\theta}}{\sqrt{Var(\widehat{\theta}_b^*)}} \tag{4.11}$$

para cada uma das B reamostras *bootstrap*. O 100α-ésimo percentil de Z_b^* é estimado por $\widehat{t}_\alpha$, tal que

$$\frac{\#\{Z_b^* \leq \widehat{t}_{(\alpha)}\}}{B} = \alpha,$$

onde $\#\{\cdot\}$ é o número de vezes que a condição é verdadeira.

Logo, o intervalo de confiança studentizado *bootstrap* é dado por

$$\left[\widehat{\theta} - t_{(1-\frac{\alpha}{2})}\sqrt{Var_B(\widehat{\theta})}\ ;\ \widehat{\theta} - t_{(\frac{\alpha}{2})}\sqrt{Var_B(\widehat{\theta})}\right]. \tag{4.12}$$

Um fato que deve ser levado em consideração na equação 4.11 é que ao calcular Z_b^*, temos que conhecer o valor de $Var(\widehat{\theta}_b^*)$. Quando não dispomos dessa quantidade, um novo esquema de reamostragem deve ser feito para estimar

o valor de $Var(\widehat{\theta}_b^*)$, para cada uma das B amostras *bootstrap*. Este esquema dar-se-á pelo duplo *bootstrap*, isto é, tomamos um número U de reamostras da b-ésima reamostra *bootstrap* e calculamos a variância das U estimativas obtidas, como descrito na equação 4.7, apenas substituindo B por U. O valor de U (que é o número de amostras *bootstrap* necessárias para a construção do intervalo de confiança *bootstrap* studentizado) é dado pelo produto de B pelo número de novas amostras *bootstrap* requerida para cada estimação de $Var(\widehat{\theta}_b^*)$. Entretanto, o elevado número requerido de reamostragens representa um grande custo computacional.

4.2.4 Intervalo *bootstrap* binomial

Considere uma população em que a probabilidade de elementos portadores de uma certa característica (sucesso) é p e a probabilidade de não ocorrência (fracasso) é $1-p$. Desta população é retirada, com reposição, todas as amostras aleatórias simples possíveis de tamanho v. Se W representa o número de sucessos obtidos nas v tentativas independentes, então a variável aleatória W tem distribuição binomial, representada por $W \sim Bin\ (v\ ,\ p)$.

Seja $\widehat{p}$ o estimador do parâmetro p. Considere a aproximação de $(\widehat{p}-p)$ por uma normal, ou seja,

$$Z = \frac{\widehat{p}-p}{\sqrt{Var(\widehat{p})}} \sim N(0\ ,\ 1). \tag{4.13}$$

Baseado na expressão 4.5 e na aproximação 4.13, um intervalo bilateral com $100(1-\alpha)\%$ de nível de confiança é dado por

$$\left[\widehat{p} - z_{(1-\frac{\alpha}{2})}\sqrt{Var(\widehat{p})}\ ;\ \widehat{p} - z_{(\frac{\alpha}{2})}\sqrt{Var(\widehat{p})}\right], \tag{4.14}$$

sendo $z_{(1-\frac{\alpha}{2})}$ e $z_{(\frac{\alpha}{2})}$ os quantis da distribuição normal padrão, e

$$Var(\widehat{p}) = \frac{\widehat{p}\,(1-\widehat{p})}{v}.$$

Para construirmos intervalos de confiança *bootstrap* a partir do intervalo apresentado na expressão 4.14, o qual também é conhecido como intervalo de Wald, calculamos a estimativa $\widehat{p}_b^*$, que é proveniente da reamostra w_b^*, e a quantidade

$$Z_b^* = \frac{\widehat{p}_b^* - \widehat{p}}{\sqrt{Var(\widehat{p}_b^*)}} \tag{4.15}$$

para cada uma das B reamostras *bootstrap*.

Logo, o intervalo de confiança binomial *bootstrap* é dado por

$$\left[\widehat{p} - z_{(1-\frac{\alpha}{2})}\sqrt{Var_B(\widehat{p})}\ ;\ \widehat{p} - z_{(\frac{\alpha}{2})}\sqrt{Var_B(\widehat{p})}\right], \tag{4.16}$$

onde

$$Var_B(\widehat{p}) = \frac{\overline{\widehat{p}^*}(1-\overline{\widehat{p}^*})}{v}.$$

Observação: Segundo Efron e Tibshirani (1993), todos os intervalos *bootstrap* discutidos neste Capítulo são intervalos aproximados, mas, em geral, os valores dessas aproximações são melhores do que os valores dos intervalos padrão.

4.3 Método para geração de uma região de confiança *bootstrap* para a curva *ROC*

A seguir, apresentamos um método para geração de uma região de confiança *bootstrap* para a curva *ROC*.

4.3.1 Método segundo Martinez

Baseado em testes diagnósticos, Martinez (2001) propõe um algoritmo para construção de uma região de confiança global para a curva *ROC* usando o método *bootstrap*.

Seja $\boldsymbol{\Upsilon} = (x_1, x_2, ..., x_n, y_1, y_2, ..., y_m)$ um vetor de dimensão $n + m$. Um teste diagnóstico que está sendo investigado, com dados contínuos, ou no mínimo categóricos ordinais, assume valores em $\boldsymbol{\Upsilon}$.

1. Consideramos Y ($Y \in \Upsilon$) e X ($X \in \Upsilon$) como sendo amostras aleatórias independentes que representam as respostas de um teste diagnóstico que está sendo investigado, para indivíduos doentes e não doentes, respectivamente. Logo, temos as amostras $\mathbf{y} = (y_1, y_2, ..., y_m)$ e $\mathbf{x} = (x_1, x_2, ..., x_n)$, onde m e n são os tamanhos amostrais para os doentes e os não doentes, respectivamente.

Sejam $\boldsymbol{\Upsilon} = (x_1, x_2, ..., x_n, y_1, y_2, ..., y_m)$ um vetor de $n + m$ observações e $\Upsilon_{(t)}$ a t-ésima estatística de ordem de $\boldsymbol{\Upsilon}$, em que $t = 1, ..., n + m$. Para uma observação Υ' ($\Upsilon' \in \boldsymbol{\Upsilon}$) e para um t fixo, a seguinte regra de decisão é considerada

$$\begin{cases} \Upsilon' \geq \Upsilon_{(t)}, \text{ se o teste é positivo,} \\ \Upsilon' < \Upsilon_{(t)}, \text{ se o teste é negativo.} \end{cases}$$

Definimos as estimativas de S_E e E_S para o teste que está sendo investigado, em um ponto de corte $\Upsilon_{(t)}$, como sendo

$$\widehat{S_E}(t) = \frac{\#\{y_{(m^{\blacktriangle})} \geq \Upsilon_{(t)}\}}{m},$$

e

$$\widehat{E_S}(t) = \frac{\#\{x_{(n^{\blacktriangle})} < \Upsilon_{(t)}\}}{n},$$

em que $\#\{\cdot\}$ é o número de vezes que a condição é verdadeira.

Logo, a curva *ROC* empírica é definida como sendo a representação gráfica contendo no eixo das abscissas $1 - \widehat{E_S}(t)$ e no eixo das ordenadas $\widehat{S_E}(t)$, para $t = 1, ..., n + m$.

2. Escolhemos B amostras *bootstrap*, onde $b = 1, 2, ..., B$, tais que $\mathbf{y}_b^* = (y_1^*, y_2^*, ..., y_m^*)$ e $\mathbf{x}_b^* = (x_1^*, x_2^*, ..., x_n^*)$ são baseadas em amostras aleatórias com reposição das observações $\mathbf{y}$ e $\mathbf{x}$, de tamanho m e n, respectivamente.

Considerando o vetor encontrado no item **1** de dimensão $n+m$ e baseado nas amostras aleatórias *bootstrap* encontradas acima, temos que $\boldsymbol{\Upsilon}_b^* = (x_1^*, x_2^*, ..., x_n^*, y_1^*, y_2^*, ..., y_m^*)$, em que $\Upsilon_{b(t)}^*$ é a t-ésima estatística de ordem de $\boldsymbol{\Upsilon}_b^*$.

Assim, as estimativas *bootstrap* de S_E e E_S são dadas por

$$\widehat{S_{Eb}}^*(t) = \frac{\#\{y_{b(m^\blacktriangle)}^* \geq \Upsilon_{b(t)}^*\}}{m},$$

e

$$\widehat{E_{Sb}}^*(t) = \frac{\#\{x_{b(n^\blacktriangle)}^* < \Upsilon_{b(t)}^*\}}{n},$$

em que $\#\{\cdot\}$ é o número de vezes que a condição é verdadeira, para $t = 1, ..., n+m$, onde $x_{b(n^\blacktriangle)}^*$, com $n^\blacktriangle = 1, 2, ..., n$, é a estatística de ordem de $\mathbf{x}_b^*$ e $y_{b(m^\blacktriangle)}^*$, com $m^\blacktriangle = 1, 2, ..., m$, é a estatística de ordem de $\mathbf{y}_b^*$.

Desta forma, obtemos B curvas *ROC bootstrap* dadas pelo gráfico de $1 - \widehat{E_{Sb}}^*(t)$ *versus* $\widehat{S_{Eb}}^*(t)$, para $t = 1, ..., n+m$ e $b = 1, 2, ..., B$.

3. Construímos um intervalo de confiança bilateral baseado em $\widehat{S_{Eb}}^*(t)$, com coeficiente de confiança de $\sqrt{(1-\alpha)}$, a cada estimativa $\widehat{S_E}(t)$ de $S_E(t)$, onde $t = 1, ..., n+m$, baseada em $\mathbf{y} = (y_1, y_2, ..., y_m)$. O intervalo de confiança em questão é baseado no método percentil e é dado por

$$\left[\widehat{L}_I(S_E(t)) \ ; \ \widehat{L}_S(S_E(t))\right] = \left[\widehat{S_{EB}}^*(t)_{\left(\frac{1-\sqrt{1-\alpha}}{2}\right)} \ ; \ \widehat{S_{EB}}^*(t)_{\left(\frac{1+\sqrt{1-\alpha}}{2}\right)}\right],$$

para $t = 1, ..., n+m$, em que $\widehat{S_{EB}}^*(t)_{(\beta)}$ é o 100β-ésimo percentil dos valores $\widehat{S_{Eb}}^*(t)$, isto é, o Bβ-ésimo valor de uma ordenação das B replicações *bootstrap* de $\widehat{S_{Eb}}^*(t)$.

4. Construímos um intervalo de confiança bilateral baseado em $\widehat{E_{Sb}}^*(t)$, com coeficiente de confiança de $\sqrt{(1-\alpha)}$, a cada estimativa $\widehat{E_S}(t)$ de $E_S(t)$, onde $t = 1, ..., n+m$, baseada em $\mathbf{x} = (x_1, x_2, ..., x_n)$. O intervalo de confiança em questão é baseado no método percentil e é dado por

$$\left[\widehat{L}_I(E_S(t)) \ ; \ \widehat{L}_S(E_S(t))\right] = \left[\widehat{E_{SB}}^*(t)_{\left(\frac{1-\sqrt{1-\alpha}}{2}\right)} \ ; \ \widehat{E_{SB}}^*(t)_{\left(\frac{1+\sqrt{1-\alpha}}{2}\right)}\right],$$

para $t = 1, ..., n+m$, em que $\widehat{E_{S_B}}^*(t)_{(\beta)}$ é o 100β-ésimo percentil dos valores $\widehat{E_{S_b}}^*(t)$, isto é, o Bβ-ésimo valor de uma ordenação das B replicações *bootstrap* de $\widehat{E_{S_b}}^*(t)$.

Entretanto, devemos lembrar que quando construímos a curva ROC, usamos o valor de $1 - E_S$, e não simplesmente o valor de E_S.

Consequentemente, o intervalo de confiança para cada estimativa $1 - \widehat{E_S}(t)$ de $1 - E_S(t)$, onde $t = 1, ..., n+m$, é dado por

$$\left[\widehat{L}_I(1 - E_S(t)) \; ; \; \widehat{L}_S(1 - E_S(t))\right] = \left[1 - \widehat{E_{S_B}}^*(t)_{\left(\frac{1+\sqrt{1-\alpha}}{2}\right)} \; ; \; 1 - \widehat{E_{S_B}}^*(t)_{\left(\frac{1-\sqrt{1-\alpha}}{2}\right)}\right].$$

5. Desta forma, obtemos a região de confiança *bootstrap* para a curva ROC, que é dada pela união das regiões de confiança criadas nos itens **3** e **4**.

Capítulo 5

Região de Confiança para a Curva *ROC*

A curva *ROC* empírica é uma estimativa para a curva *ROC* teórica. Por ser uma curva amostral, ela carrega consigo uma incerteza. Neste Capítulo, vamos fazer uma revisão de alguns métodos utilizados para determinar a incerteza associada à estimativa da curva *ROC* e propomos dois novos métodos para a estimação desta incerteza. Essa incerteza será determinada por uma região probabilística, que deve conter a curva *ROC* teórica com uma probabilidade previamente conhecida.

Em geral, a região probabilística (chamada também de região de probabilidade, ou de região de confiança, ou de região de incerteza) associada a uma curva *ROC* é determinada graficamente por duas curvas limites, uma superior e outra inferior, que delimitam a região em questão.

Banda de confiança é uma parte da região de confiança associada a uma curva *ROC* juntamente com essas curvas limites (fronteiras). Temos a banda de confiança inferior, que representa a parte da região de probabilidade que vai desde a fronteira inferior até esta curva *ROC*, e a banda de confiança superior, que representa a parte da região de probabilidade que vai desta curva *ROC* até a fronteira superior. Quando nos referimos a bandas de confiança, estamos tratando da região de confiança em seu todo, que vem a ser a união da banda de confiança

superior com a banda de confiança inferior.

Definimos três tipos de regiões de confiança para a curva ROC: as pontuais, descritas através dos métodos das médias verticais, médias otimizadas, médias limiares e médias limiares otimizadas, as regionais, descritas pelo método das regiões de confiança de juntas simultâneas e as globais, descritas através do método das bandas de largura fixa.

Apresentamos a seguir, alguns dos métodos utilizados na construção dessa região probabilística.

5.1 Bandas de confiança pontuais

Essas bandas de confiança são compostas por um conjunto de intervalos de confiança aplicados a cada ponto de corte estimado da curva ROC.

São considerados os intervalos para os pontos de FP associados aos respectivos valores de VP, ou intervalos para os pontos de VP associados aos respectivos valores de FP, ou até mesmo regiões bidimensionais para os pares de $(FP\ ,\ VP)$ representados na curva ROC.

Hilgers (1991) propõe um método livre de distribuição, que calcula, separadamente, intervalos de confiança para VP e FP, e posteriormente faz uma combinação dos mesmos. Schäfer (1994) denominou a proposta de Hilgers de "método de dois estágios", argumentando que os intervalos de confiança são amplos e produzem uma região com probabilidade de cobertura superior que a desejada.

Segundo Martinez *et al.* (2003), é útil encontrar região de confiança baseada em bandas pontuais quando o objetivo é a visualização da variabilidade amostral de VP e FP nas coordenadas da curva ROC correspondentes aos pontos de corte empíricos, mas não na curva ROC em sua totalidade.

Tal região de probabilidade é descrita pelos métodos das: médias verticais, médias otimizadas, médias limiares e médias limiares otimizadas.

5.1.1 Método das médias verticais

O método das médias verticais também é conhecido como *vertical averaging* (**VA**).

Para um dado valor de FP_i ($0 \leq FP_i \leq 1$), no qual $FP_i = (i-1)l$, temos que existe um único $i = 1, ..., 1 + \frac{1}{l}$, onde l ($l > 0$) é o comprimento (fixo) entre FP_i e FP_{i+1} (ou FP_i e FP_{i-1}), isto é, $|FP_{i+1} - FP_i| = l$ (ou $|FP_i - FP_{i-1}| = l$), de tal forma que $FP_{i-1} < FP_i < FP_{i+1}$.

Usando interpolação linear, vamos encontrar o valor de VP_i, para um dado valor de FP_i, da seguinte forma

$$VP_i = \begin{cases} VP_{i-1} + \dfrac{(FP_i - FP_{i-1})(VP_{i+1} - VP_{i-1})}{(FP_{i+1} - FP_{i-1})} & , se\ 0 \leq FP_i < 1, \\ 1 & , se\ FP_i = 1. \end{cases} \tag{5.1}$$

Definimos as funções

$$f : [0,1] \mapsto [0,1],$$

e

$$f_j : [0,1] \mapsto [0,1],$$

onde f associa, no conjunto dos dados originais, um valor de VP_i para um dado valor de FP_i e f_j associa, nas j-ésimas reamostragens *bootstrap*, onde $j = 1, ..., B$, um valor de VP_{ij} para um dado valor de FP_i.

A partir de uma curva ROC estimada, geramos B curvas ROC pelo método de reamostragem *bootstrap*.

Baseado em Macskassy *et al.* (2003) e Macskassy e Provost (2004), calculamos bandas pontuais pelo método VA da seguinte maneira: percorremos o eixo das abscissas de $FP_i = 0$ até $FP_i = 1$ em um *grid* de valores para FP_i, de comprimento l, de tal forma que $FP_i = (i-1)l$, para $i = 1, ..., 1 + \frac{1}{l}$. Para cada valor de FP_i, onde $i = 1, ..., 1 + \frac{1}{l}$, determinamos o respectivo valor de $\overline{VP}_i$,

como sendo

$$\overline{VP}_i = \frac{1}{B}\sum_{j=1}^{B} VP_{ij}, \tag{5.2}$$

no qual

$$VP_{ij} = f_j(FP_i), \tag{5.3}$$

para $j = 1, ..., B$. Logo, $\overline{VP}_i$ é a média, em relação às diferentes reamostragens, dos valores de VP_{ij}. Consequentemente, obtemos pontos de corte da forma $(FP_i\ ,\ \overline{VP}_i)$, para os diferentes valores de i, também conhecidos como pontos de corte médio. Uma curva ROC média é traçada unindo-se todos os pontos de corte médios estimados durante esse procedimento.

Em cada FP_i, para $i = 1, ..., 1 + \frac{1}{l}$, baseado na amostra *bootstrap* $\{VP_{ij}, j = 1, ..., B\}$, determinamos por um dos métodos assumidos no Capítulo 4, um intervalo vertical centrado em $\overline{VP}_i$, como sendo

$$[VP_i''\ ;\ VP_i'].$$

Logo, $|VP_i' - \overline{VP}_i| = |\overline{VP}_i - VP_i''|$.

Após a geração desses intervalos verticais em cada ponto de corte $(FP_i\ ,\ \overline{VP}_i)$, para $i = 1, ..., 1 + \frac{1}{l}$, bandas de confiança superior e inferior são encontradas. A banda superior corresponde à região que varia desde a curva ROC média até a curva limite superior, dada quando todos os limites superiores de cada intervalo são unidos, isto é, quando unimos todos os pontos $(FP_i\ ,\ VP_i')$, para os diferentes valores de i. A banda inferior corresponde à região que varia desde a curva limite inferior, dada quando todos os limites inferiores de cada intervalo são unidos, isto é, quando unimos todos os pontos $(FP_i\ ,\ VP_i'')$, para os diferentes valores de i, até a curva ROC média. Logo, tendo as bandas de confiança superior e inferior, temos a região de confiança para a curva ROC média. Esperamos que essa região de confiança tenha uma probabilidade de cobertura de $(1 - \alpha)$ do total das curvas ROC geradas pelo *bootstrap*, onde α $(0 \leq \alpha \leq 1)$ é o nível de significância.

A Figura 5.1 ilustra essa metodologia, em que foi gerada uma região de probabilidade para a curva ROC utilizando o método das médias verticais.

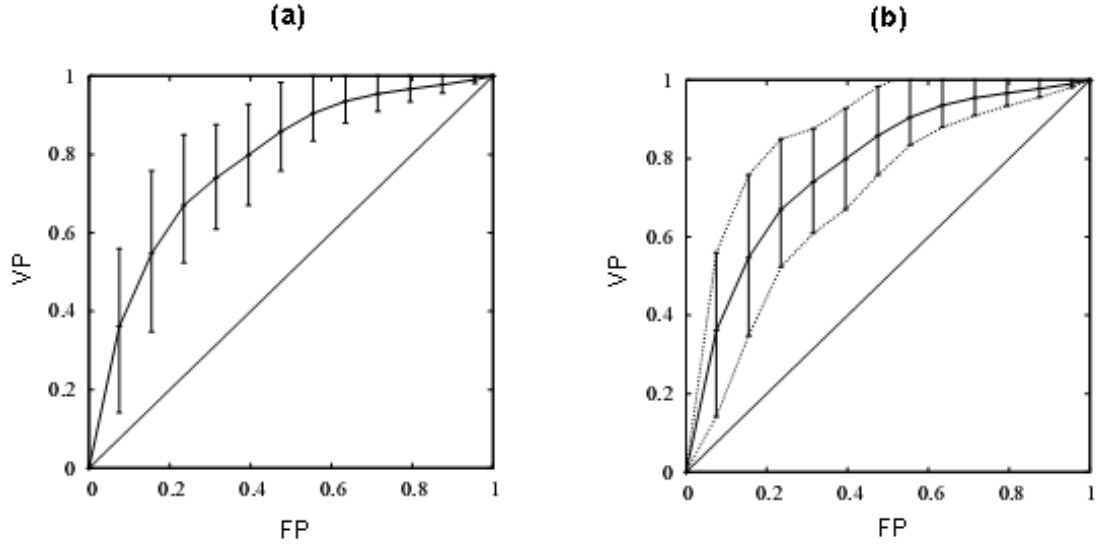

FIGURA 5.1: Ilustração do método VA, onde (a) representa a forma de geração dos intervalos de confiança em torno dos pontos de corte (FP_i , $\overline{VP}_i$) e (b) representa a região de incerteza relacionada com esses intervalos.

5.1.2 Método das médias otimizadas

Macskassy e Provost (2004) relatam que existem vários métodos para se trabalhar com dois intervalos de confiança unidimensionais ao mesmo tempo. Propomos um novo método para geração de regiões de confiança pontuais para a curva *ROC*, o qual chamaremos de método das médias otimizadas (**MO**).

Para um dado valor de FP_i, onde $i = 1, ..., 1 + \frac{1}{l}$, encontramos o respectivo valor de VP_i, usando interpolação linear, a partir da equação 5.1.

Para um dado valor de VP_i $(0 \leq VP_i \leq 1)$, temos que existe um único $i = 1, ..., 1 + \frac{1}{l}$, onde $FP_i = (i-1)l$ e $|FP_{i+1} - FP_i| = l = |FP_i - FP_{i-1}|$, para l fixo $(l > 0)$, de tal forma que $VP_{i-1} < VP_i < VP_{i+1}$.

Usando interpolação linear, vamos encontrar o valor de FP_i, para um dado valor de VP_i, da seguinte forma

$$FP_i = \begin{cases} FP_{i-1} + \dfrac{(VP_i - VP_{i-1})(FP_{i+1} - FP_{i-1})}{(VP_{i+1} - VP_{i-1})} & , se\ 0 < VP_i \leq 1, \\ 0 & , se\ VP_i = 0. \end{cases} \quad (5.4)$$

Definimos as funções

$$g : [0,1] \mapsto [0,1],$$

e

$$g_j : [0,1] \mapsto [0,1],$$

onde g associa, no conjunto dos dados originais, um valor de FP_i para um dado valor de VP_i e g_j associa, nas j-ésimas reamostragens *bootstrap*, onde $j = 1, ..., B$, um valor de FP_{ij} para um dado valor de VP_i.

A partir de uma curva ROC estimada, geramos B curvas ROC pelo método de reamostragem *bootstrap*.

Propomos a seguinte forma de calcular bandas pontuais pelo método MO: percorremos o eixo das abscissas de $FP_i = 0$ até $FP_i = 1$ em um *grid* de valores para FP_i, de comprimento l ($l > 0$), de tal forma que $FP_i = (i-1)l$, para $i = 1, ..., 1 + \frac{1}{l}$. Para cada valor de FP_i, onde $i = 1, ..., 1 + \frac{1}{l}$, determinamos o respectivo valor de $\overline{VP}_i$, como mostrado na equação 5.2. Logo, obtemos pontos de corte da forma (FP_i , $\overline{VP}_i$), para os diferentes valores de i.

Em cada FP_i, para $i = 1, ..., 1 + \frac{1}{l}$, baseado na amostra *bootstrap* $\{VP_{ij}, j = 1, ..., B\}$, determinamos por um dos métodos assumidos no Capítulo 4, um intervalo vertical centrado em $\overline{VP}_i$, como sendo

$$[VP_{i_{(II)}} \ ; \ VP_{i_{(I)}}].$$

Logo, $|VP_{i_{(I)}} - \overline{VP}_i| = |\overline{VP}_i - VP_{i_{(II)}}|$.

Posteriormente, fixamos os diferentes valores de $\overline{VP}_i$, para $i = 1, ..., 1+\frac{1}{l}$, e percorremos o eixo das ordenadas de $VP_i = 0$ até $VP_i = 1$ em intervalos não regulares[1]. Para cada valor de $\overline{VP}_i$, onde $i = 1, ..., 1 + \frac{1}{l}$, determinamos o respectivo valor de $\overline{FP}_i$, como sendo

$$\overline{FP}_i = \frac{1}{B} \sum_{j=1}^{B} FP_{ij}, \tag{5.5}$$

[1]: *Os intervalos não tem o mesmo comprimento.*

no qual

$$FP_{ij} = g_j(VP_i), \tag{5.6}$$

para $j = 1, ..., B$. Logo, $\overline{FP}_i$ é a média, em relação às diferentes reamostragens, dos valores de FP_{ij}.

Em cada $\overline{VP}_i$, para $i = 1, ..., 1 + \frac{1}{l}$, baseado na amostra *bootstrap* $\{FP_{ij}, j = 1, ..., B\}$, determinamos por um dos métodos assumidos no Capítulo 4, um intervalo horizontal centrado em $\overline{FP}_i$, como sendo

$$[FP_{i_{(II)}} \ ; \ FP_{i_{(I)}}].$$

Logo, $|FP_{i_{(I)}} - \overline{FP}_i| = |\overline{FP}_i - FP_{i_{(II)}}|$.

Assim, para cada valor de i, onde $i = 1, ..., 1 + \frac{1}{l}$, o método MO gera um intervalo vertical $[VP_{i_{(II)}} \ ; \ VP_{i_{(I)}}]$ e um intervalo horizontal $[FP_{i_{(II)}} \ ; \ FP_{i_{(I)}}]$. Posteriormente, fazemos combinações desses intervalos para encontrarmos as fronteiras superior e inferior que delimitam a região de incerteza para a curva *ROC*. A Figura 5.2 ilustra essas combinações.

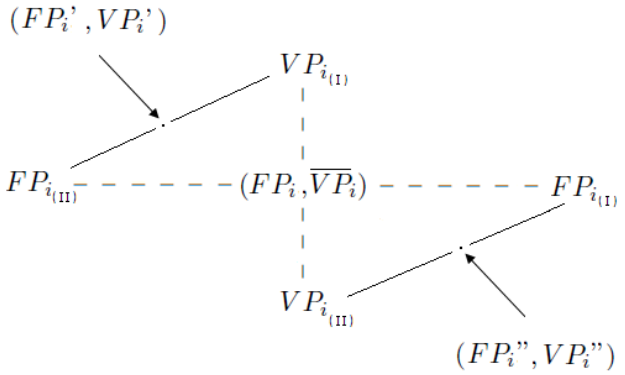

FIGURA 5.2: Combinações efetuadas pelo método MO em torno do ponto de corte $(FP_i \ , \ \overline{VP}_i)$.

Da Figura 5.2, notamos que os pontos que correspondem à fronteira superior, denotada por $(FP_i' \ , \ VP_i')$, são encontrados quando se calcula a média entre os limites superiores dos intervalos verticais e os limites inferiores dos intervalos horizontais, para $i = 1, ..., 1 + \frac{1}{l}$. Notamos também, que os pontos que correspondem à fronteira inferior, denotada por $(FP_i'' \ , \ VP_i'')$, são encontrados

quando se calcula a média entre os limites inferiores dos intervalos verticais e os limites superiores dos intervalos horizontais, para $i = 1, ..., 1 + \frac{1}{l}$.

Conhecidas as formas de geração dos pontos pertencentes às fronteiras superior e inferior, encontramos as bandas de confiança superior e inferior para a curva *ROC*. A banda superior corresponde à região que varia desde a curva *ROC* até a curva limite superior, dada quando unimos, para $i = 1, ..., 1 + \frac{1}{l}$, os valores

$$\left(\frac{FP_i + FP_{i_{(II)}}}{2} \ , \ \frac{\overline{VP}_i + VP_{i_{(I)}}}{2} \right).$$

A banda inferior corresponde à região que varia desde a curva limite inferior, dada quando unimos, para $i = 1, ..., 1 + \frac{1}{l}$, os valores

$$\left(\frac{FP_i + FP_{i_{(I)}}}{2} \ , \ \frac{\overline{VP}_i + VP_{i_{(II)}}}{2} \right)$$

até a curva *ROC*. Logo, tendo as bandas de confiança superior e inferior, temos a região de confiança para a curva *ROC*. Esperamos que essa região de confiança tenha uma probabilidade de cobertura de $(1-\alpha)$ do total das curvas *ROC* geradas pelo *bootstrap*, onde α $(0 \leq \alpha \leq 1)$ é o nível de significância.

A Figura 5.3 ilustra essa metodologia, em que foi gerada uma região de confiança para a curva *ROC* utilizando o método das médias otimizadas.

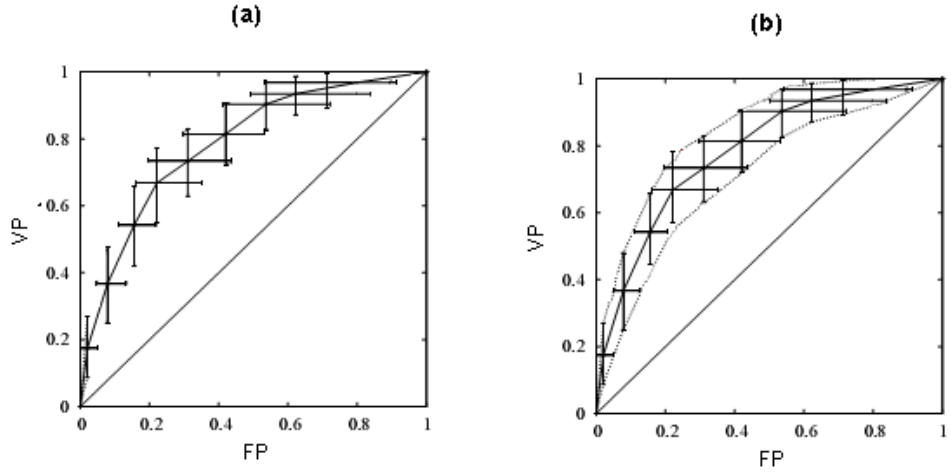

FIGURA 5.3: Ilustração do método MO, onde (a) representa a forma de geração dos intervalos de confiança em torno dos pontos de corte $(FP_i \ , \ \overline{VP}_i)$ e (b) representa a região de incerteza relacionada com esses intervalos.

5.1.3 Método das médias limiares

O método das médias limiares também é conhecido como *threshold averaging* (**TA**).

Considere uma quantidade finita de pontos de corte representados por $(FP_r\ ,\ VP_r)$, onde $r = 1, 2, ..., R + 1$, sendo que R são os possíveis resultados de um teste diagnóstico.

Definimos as funções

$$f : [0, 1] \mapsto [0, 1],$$

e

$$f_j : [0, 1] \mapsto [0, 1],$$

onde f associa, no conjunto dos dados originais, um valor de VP_r para um dado valor de FP_r e f_j associa, nas j-ésimas reamostragens *bootstrap*, onde $j = 1, ..., B$, um valor de VP_{rj} para um dado valor de FP_r.

Baseado em Fawcett (2003) e Macskassy e Provost (2004), calculamos bandas pontuais pelo método TA da seguinte maneira: reamostramos o conjunto de dados uma primeira vez e obtemos os pontos de corte $(FP_{r1}\ ,\ VP_{r1})$, onde $r = 1, ..., R + 1$, de tal forma que eles sejam ordenados da seguinte maneira: $(0\ ,\ 0) = (FP_{(R+1)1}\ ,\ VP_{(R+1)1}) < ... < (FP_{11}\ ,\ VP_{11}) = (1\ ,\ 1)$. Seguindo este mesmo raciocínio, quando realizamos a B-ésima reamostragem *bootstrap*, obtemos os pontos de corte $(FP_{rB}\ ,\ VP_{rB})$, onde $r = 1, ..., R+1$, de tal forma que, quando ordenados, estes são representados por: $(0\ ,\ 0) = (FP_{(R+1)B}\ ,\ VP_{(R+1)B}) < ... < (FP_{1B}\ ,\ VP_{1B}) = (1\ ,\ 1)$. Assim, criamos um *array* dos pontos de corte que, quando ordenados, formam um conjunto de limiares. A Figura 5.4 ilustra estas ordenações durante uma j-ésima reamostragem *bootstrap* qualquer.

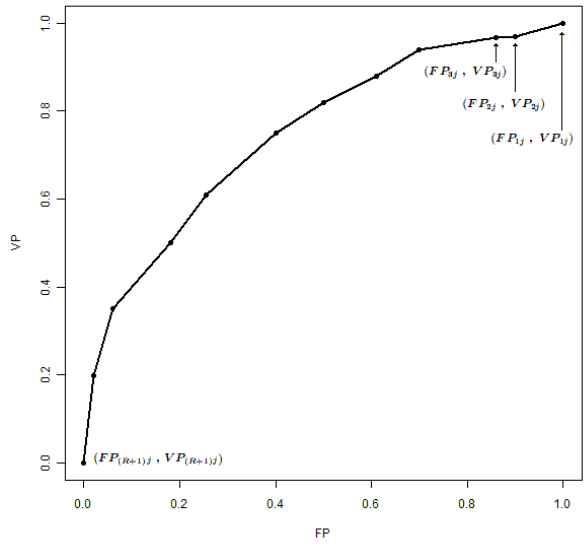

FIGURA 5.4: Ordenação dos pontos gerados durante a j-ésima reamostragem.

Logo, encontramos uma "nuvem" de pontos ao redor dos pontos de corte que serviram para traçar a curva ROC inicial, ou seja, para **cada** ponto de corte inicialmente encontrado $(FP_r\ ,\ VP_r)$, após realizadas todas as B reamostragens *bootstrap*, associamos os pontos $(FP_{rj}\ ,\ VP_{rj})$, no qual $r = 1, ..., R+1$ e $j = 1, ..., B$.

Posteriormente, para cada valor do ponto $(FP_r\ ,\ VP_r)$ inicialmente encontrado, onde $r = 1, ..., R+1$, determinamos o respectivo valor de $\overline{VP}_r$, como sendo

$$\overline{VP}_r = \frac{1}{B}\sum_{j=1}^{B} VP_{rj}, \tag{5.7}$$

no qual

$$VP_{rj} = f_j(FP_r), \tag{5.8}$$

para $j = 1, ..., B$. Logo, $\overline{VP}_r$ é a média, em relação às diferentes reamostragens, dos valores de VP_{rj} para o ponto $(FP_r\ ,\ VP_r)$, nos diferentes valores de r. Consequentemente, obtemos pontos de corte da forma $(FP_r\ ,\ \overline{VP}_r)$, também conhecidos como pontos de corte médio, para os diferentes valores de r. Uma curva ROC média é traçada unindo-se todos os pontos de corte médios estimados durante esse procedimento.

Em cada ponto $(FP_r\ ,\ \overline{VP}_r)$, para $r = 1, ..., R+1$, baseado na amostra

bootstrap $\{VP_{rj}, j = 1, ..., B\}$, determinamos por um dos métodos assumidos no Capítulo 4, um intervalo vertical centrado em $\overline{VP}_r$, como sendo

$$[VP''_r \ ; \ VP'_r].$$

Logo, $|VP'_r - \overline{VP}_r| = |\overline{VP}_r - VP''_r|$.

Após a geração desses intervalos verticais em cada ponto de corte médio $(FP_r \ , \ \overline{VP}_r)$, para $r = 1, ..., R + 1$, bandas de confiança superior e inferior são encontradas. A banda superior corresponde à região que varia desde a curva *ROC* média até a curva limite superior, dada quando todos os limites superiores de cada intervalo são unidos, isto é, quando unimos todos os pontos $(FP_r \ , \ VP'_r)$, para os diferentes valores de r. A banda inferior corresponde à região que varia desde a curva limite inferior, dada quando todos os limites inferiores de cada intervalo são unidos, isto é, quando unimos todos os pontos $(FP_r \ , \ VP''_r)$, para os diferentes valores de r, até a curva *ROC* média. Logo, tendo as bandas de confiança superior e inferior, temos a região de confiança para a curva *ROC* média. Esperamos que essa região de confiança tenha uma probabilidade de cobertura de $(1 - \alpha)$ do total das curvas *ROC* geradas pelo *bootstrap*, onde α $(0 \leq \alpha \leq 1)$ é o nível de significância.

A Figura 5.5 ilustra essa metodologia, em que foi gerada uma região de probabilidade para a curva *ROC* utilizando o método das médias limiares.

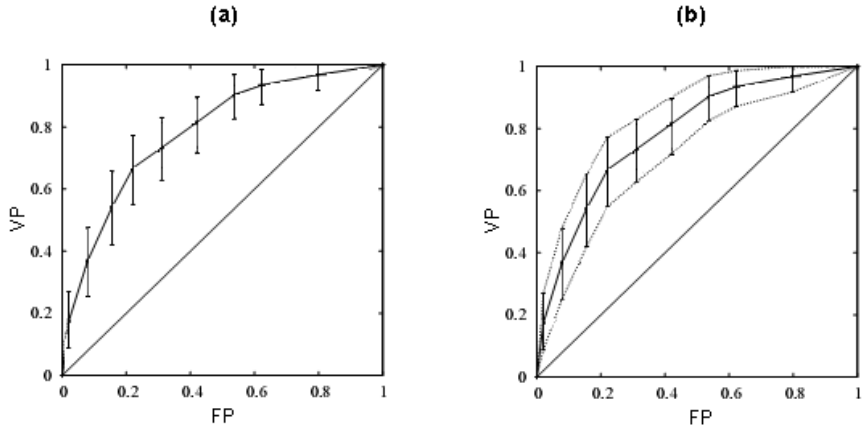

FIGURA 5.5: Ilustração do método TA, onde (a) representa a forma de geração dos intervalos de confiança em torno dos pontos de corte $(FP_r \ , \ \overline{VP}_r)$ e (b) representa a região de incerteza relacionada com esses intervalos.

Observação:- No método TA, diferentemente do método VA, só nos interessa os limiares do modelo. Assim, no método TA não estamos interessados em encontrar intervalos de confiança verticais nos pontos de corte que não são fornecidos diretamente dos possíveis resultados do teste diagnóstico.

5.1.4 Método das médias limiares otimizadas

Propomos um outro método para geração de regiões de confiança pontuais para a curva *ROC*, o qual chamaremos de método das médias limiares otimizadas (**MLO**).

Definimos as funções

$$g : [0,1] \mapsto [0,1],$$

e

$$g_j : [0,1] \mapsto [0,1],$$

onde g associa, no conjunto dos dados originais, um valor de FP_r para um dado valor de VP_r e g_j associa, nas j-ésimas reamostragens *bootstrap*, onde $j = 1, ..., B$, um valor de FP_{rj} para um dado valor de VP_r.

Propomos a seguinte forma de calcular bandas pontuais pelo método MLO: reamostramos o conjunto de dados uma primeira vez e obtemos os pontos de corte $(FP_{r1} \ , \ VP_{r1})$, onde $r = 1, ..., R+1$, de tal forma que eles sejam ordenados da seguinte maneira: $(0 \ , \ 0) = (FP_{(R+1)1} \ , \ VP_{(R+1)1}) < ... < (FP_{11} \ , \ VP_{11}) = (1 \ , \ 1)$. Seguindo este mesmo raciocínio, quando realizamos a B-ésima reamostragem *bootstrap*, obtemos os pontos de corte $(FP_{rB} \ , \ VP_{rB})$, onde $r = 1, ..., R+1$, de tal forma que, quando ordenados, estes são representados por: $(0 \ , \ 0) = (FP_{(R+1)B} \ , \ VP_{(R+1)B}) < ... < (FP_{1B} \ , \ VP_{1B}) = (1 \ , \ 1)$. Assim, criamos um *array* dos pontos de corte que, quando ordenados, formam um conjunto de limiares.

Desta forma, encontramos uma "nuvem" de pontos ao redor dos pontos

de corte que serviram para traçar a curva ROC inicial, ou seja, para cada ponto de corte inicialmente encontrado $(FP_r\ ,\ VP_r)$, após realizadas todas as B reamostragens *bootstrap*, associamos os pontos $(FP_{rj}\ ,\ VP_{rj})$, no qual $r = 1, ..., R+1$ e $j = 1, ..., B$.

Para cada valor do ponto $(FP_r\ ,\ VP_r)$ inicialmente encontrado, onde $r = 1, ..., R+1$, determinamos o respectivo valor de $\overline{VP}_r$, como mostrado na equação 5.7. Logo, obtemos pontos da forma $(FP_r\ ,\ \overline{VP}_r)$, para os diferentes valores de r.

Em cada ponto $(FP_r\ ,\ \overline{VP}_r)$, para $r = 1, ..., R+1$, baseado na amostra *bootstrap* $\{VP_{rj}, j = 1, ..., B\}$, determinamos por um dos métodos assumidos no Capítulo 4, um intervalo vertical centrado em $\overline{VP}_r$, como sendo

$$[VP_{r_{(II)}}\ ;\ VP_{r_{(I)}}].$$

Logo, $|VP_{r_{(I)}} - \overline{VP}_r| = |\overline{VP}_r - VP_{r_{(II)}}|$.

Para cada valor do ponto $(FP_r\ ,\ VP_r)$ inicialmente encontrado, onde $r = 1, ..., R+1$, determinamos o respectivo valor de $\overline{FP}_r$, como sendo

$$\overline{FP}_r = \frac{1}{B}\sum_{j=1}^{B} FP_{rj}, \tag{5.9}$$

no qual

$$FP_{rj} = g_j(VP_r), \tag{5.10}$$

para $j = 1, ..., B$. Logo, $\overline{FP}_r$ é a média, em relação às diferentes reamostragens, dos valores de FP_{rj} para o ponto $(FP_r\ ,\ VP_r)$, nos diferentes valores de r. Consequentemente, obtemos pontos da forma $(\overline{FP}_r\ ,\ VP_r)$, para os diferentes valores de r.

Em cada ponto $(\overline{FP}_r\ ,\ VP_r)$, para $r = 1, ..., R+1$, baseado na amostra *bootstrap* $\{FP_{rj}, j = 1, ..., B\}$, determinamos por um dos métodos assumidos no Capítulo 4, um intervalo horizontal centrado em $\overline{FP}_r$, como sendo

$$[FP_{r_{(II)}}\ ;\ FP_{r_{(I)}}].$$

Logo, $|FP_{r_{(I)}} - \overline{FP}_r| = |\overline{FP}_r - FP_{r_{(II)}}|$.

Assim, usando inicialmente o ponto $(FP_r\ ,\ VP_r)$, onde $r = 1, ..., R+1$, o método MLO gera um intervalo vertical $[VP_{r_{(II)}}\ ;\ VP_{r_{(I)}}]$ e um intervalo horizontal $[FP_{r_{(II)}}\ ;\ FP_{r_{(I)}}]$ centrados, respectivamente, nos pontos $(FP_r\ ,\ \overline{VP}_r)$ e $(\overline{FP}_r\ ,\ VP_r)$. Na Figura 5.6, ilustramos a forma como este método encontra os pontos que servirão para formar as bandas de confiança superior e inferior.

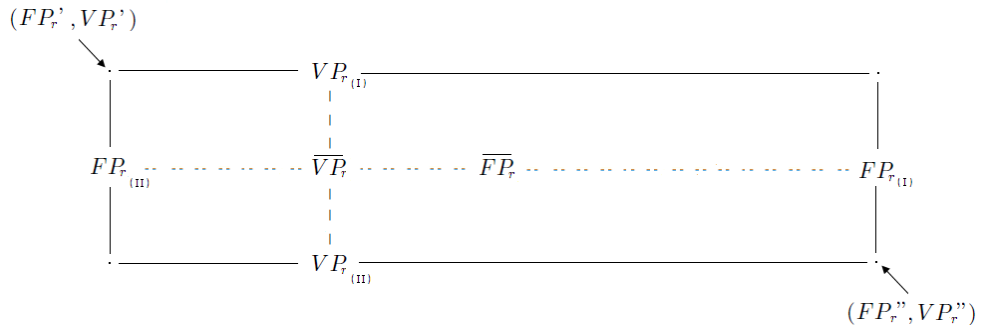

FIGURA 5.6: Pontos que formam as fronteiras de confiança encontradas pelo método MLO.

Da Figura 5.6, notamos que os pontos que correspondem à fronteira superior, denotada por $(FP'_r\ ,\ VP'_r)$, são encontrados quando unimos os cantos superiores esquerdos dos "retângulos" formados pelos intervalos verticais e pelos intervalos horizontais ao redor do ponto de corte inicialmente estimado $(FP_r\ ,\ VP_r)$, para $r = 1, ..., R+1$. Notamos também, que os pontos que correspondem à fronteira inferior, denotada por $(FP''_r\ ,\ VP''_r)$, são encontrados quando unimos os cantos inferiores esquerdos dos "retângulos" formados pelos intervalos verticais e pelos intervalos horizontais ao redor do ponto de corte inicialmente estimado $(FP_r\ ,\ VP_r)$, para $r = 1, ..., R+1$.

Conhecidas as formas de geração dos pontos pertencentes às fronteiras superior e inferior, encontramos as bandas de confiança superior e inferior para a curva ROC. A banda superior corresponde à região que varia desde a curva ROC inicialmente estimada até a curva limite superior, dada quando unimos, para $r = 1, ..., R+1$, os valores

$$\left(FP_{r_{(II)}}\ ,\ VP_{r_{(I)}}\right).$$

A banda inferior corresponde à região que varia desde a curva limite inferior, dada quando unimos, para $r = 1, ..., R + 1$, os valores

$$\left(FP_{r_{(I)}} \ , \ VP_{r_{(II)}} \right),$$

até a curva *ROC* inicialmente estimada. Logo, tendo as bandas de confiança superior e inferior, temos a região de confiança para a curva *ROC*. Esperamos que essa região de confiança tenha uma probabilidade de cobertura de $(1 - \alpha)$ do total das curvas *ROC* geradas pelo *bootstrap*, onde α $(0 \leq \alpha \leq 1)$ é o nível de significância.

A Figura 5.7 ilustra essa metodologia, em que foi gerada uma região de confiança para a curva *ROC* utilizando o método das médias limiares otimizadas.

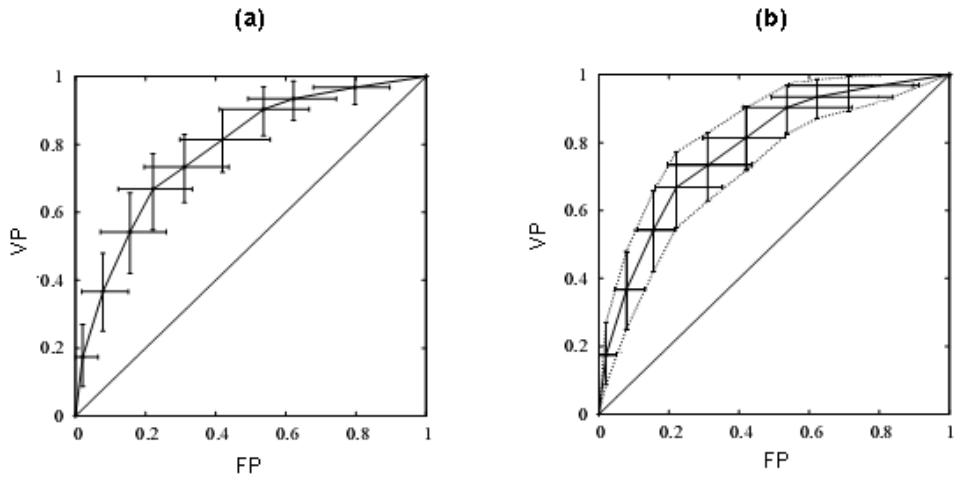

FIGURA 5.7: Ilustração do método MLO, onde (a) representa a forma de geração dos intervalos de confiança nas proximidades dos pontos de corte inicialmente estimados $(FP_r \ , \ VP_r)$ e (b) representa a região de incerteza relacionada com esses intervalos.

5.2 Bandas de confiança regionais

Em determinadas circunstâncias, decisões são tomadas somente em uma parte específica da curva *ROC*. Essas decisões podem ser baseadas em um intervalo restrito para FP ou para VP, conforme o interesse do pesquisador. Quando situações desta forma ocorrem, a estimação de bandas de confiança regionais são as mais indicadas.

Jensen *et al.* (2000) propõem um método pelo qual é produzida uma região de confiança mais estreita que a estimada pelos métodos das bandas globais, porém dirigida a uma região restrita da curva *ROC*.

Campbell (1994) propõe um método para estimação das bandas de confiança regionais baseado na estatística de *Kolmogorov-Smirnov* e outro baseado no *bootstrap*.

Tais bandas são descritas pelo método das regiões de confiança de juntas simultâneas.

5.2.1 Método das regiões de confiança de juntas simultâneas

O método das regiões de confiança de juntas simultâneas também é conhecido como *simultaneous joint confidence regions* (**SJR**). Informalmente, ele pode ser chamado de "método do retângulo local".

O método SJR gera bandas de confiança baseadas no teste de *Kolmogorov-Smirnov* (KS), para identificar um intervalo de confiança regional para VP e FP, independentemente (ver Apêndice B).

Baseado em Campbell (1994) e Macskassy e Provost (2004), calculamos bandas regionais pelo método SJR da seguinte maneira: percorremos o eixo das abscissas de $FP_i = 0$ até $FP_i = 1$ em um *grid* de valores para FP_i, de comprimento l ($l > 0$), de tal forma que $FP_i = (i-1)l$, para $i = 1, ..., 1 + \frac{1}{l}$, e encontramos o respectivo valor VP_i, para um dado valor de FP_i, usando interpolação linear, a partir da equação 5.1. Logo, para os diferentes valores de i, temos os pontos de corte (FP_i , VP_i).

Usamos, para criação deste tipo de região probabilística, intervalos centrados apenas em uma curva ROC estimada e, a partir desta curva, geramos B curvas ROC pelo processo de reamostragem *bootstrap*.

Sejam g uma distância horizontal qualquer ao longo de cada um dos valores FP_i e h uma distância vertical qualquer ao longo de cada um dos valores VP_i, onde $i = 1, ..., 1 + \frac{1}{l}$, com um nível de confiança de $(1-\alpha)$, onde $0 \leq \alpha \leq 1$,

para cada um desses dois valores.

Para um ponto de corte qualquer estimado, o método SJR gera, em torno deste ponto, uma região representada por um retângulo da forma $(FP_i \pm g \ , \ VP_i \pm h)$, para os diferentes valores de i, com um nível de confiança de $(1-\alpha)^2$. Assim, em torno de cada ponto de corte $(FP_i \ , \ VP_i)$, temos um retângulo com comprimento $2g$ e com altura $2h$. Na Figura 5.8, ilustramos este procedimento.

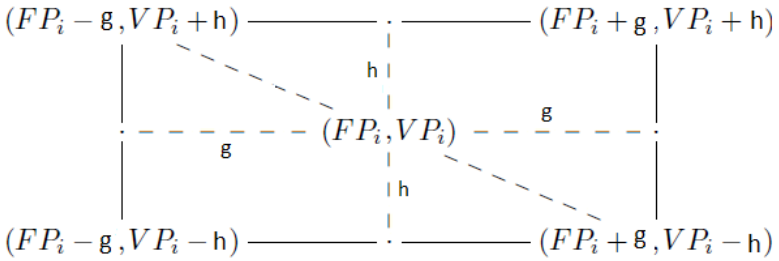

FIGURA 5.8: Retângulo em torno do ponto de corte $(FP_i \ , \ VP_i)$.

Após a geração desses retângulos, bandas de confiança superior e inferior são encontradas. A banda superior corresponde à região que varia desde a curva ROC inicialmente estimada até a curva limite superior, dada quando unimos os cantos superiores esquerdos dos retângulos, ou seja, quando unimos todos os valores $(FP_i - g \ , \ VP_i + h)$, para os diferentes valores de i. A banda inferior corresponde à região que varia desde a curva limite inferior, dada quando unimos os cantos inferiores direitos dos retângulos, ou seja, quando unimos todos os valores $(FP_i + g \ , \ VP_i - h)$, para os diferentes valores de i, até a curva ROC inicialmente estimada. Logo, tendo as bandas de confiança superior e inferior, temos a região de confiança para a curva ROC. Assim, a região de confiança para esta curva ROC deve conter retângulos de mesma dimensão e estes, por sua vez, devem conter sempre o ponto $(FP_i \ , \ VP_i)$ com uma probabilidade de $(1-\alpha)^2$, onde α $(0 \leq \alpha \leq 1)$ é o nível de significância.

A estatística de *Kolmogorov-Smirnov* justifica o uso destes retângulos para geração da região de confiança regional para a curva ROC. Assim, buscamos valores para as distâncias g e h máximos para formação de retângulos que contenham o ponto de corte $(FP_i \ , \ VP_i)$ com um nível de confiança de $(1-\alpha)^2$.

A Figura 5.9 ilustra essa metodologia, em que foi gerada uma região de probabilidade para a curva *ROC* a partir do método das regiões de confiança de juntas simultâneas.

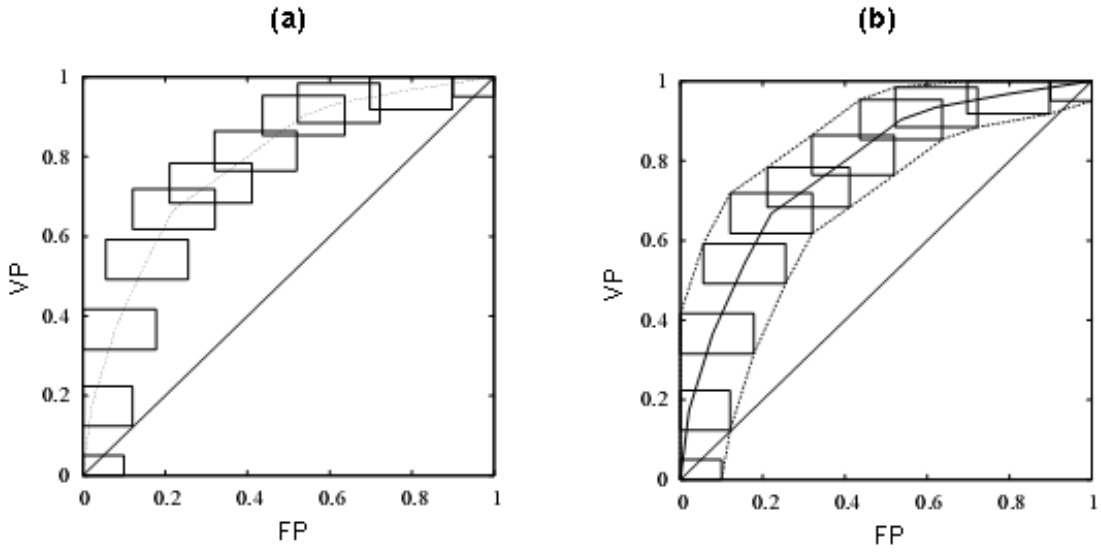

FIGURA 5.9: Ilustração do método SJR, onde (a) representa a forma de geração dos intervalos de confiança em torno dos pontos de corte (FP_i , VP_i) e (b) representa a região de incerteza relacionada com esses intervalos.

Observação:- Macskassy e Provost (2004), diferentemente de Campbell (1994), dizem que os retângulos devem conter os pontos de corte (FP_i , VP_i) com uma probabilidade de cobertura de $(1-\alpha)$, onde α $(0 \leq \alpha \leq 1)$ é o nível de significância, e não de $(1-\alpha)^2$ como seria esperado, pois estudos empíricos mostram que uma confiança de $(1-\alpha)$ é mais precisa do que uma confiança teórica de $(1-\alpha)^2$, o qual é mais fraca.

5.3 Bandas de confiança globais

O estudo de bandas globais é útil quando são adotadas regras de decisão que não são baseadas em uma região específica da curva *ROC* (não tem pontos de corte específicos), mas sim quando o interesse é a visualização do desempenho global do que está sendo investigado.

Tais bandas são descritas pelo método das bandas de largura fixa.

5.3.1 Método das bandas de largura fixa

O método das bandas de largura fixa também é conhecido como *fixed-width bands* (**FWB**).

Qualquer tipo de região de confiança que cubra completamente a curva ROC estimada, com uma alta probabilidade, pode ser usada.

Campbell (1994) propõe a formação de bandas de confiança globais com um ângulo ($\varrho^{'}$) entre a razão dos desvios padrões associados aos grupos dos indivíduos doentes e não doentes, respectivamente, dado por

$$\begin{aligned}\varrho^{'} &= \frac{\sqrt{\frac{S_E\ (1-S_E)}{m}}}{\sqrt{\frac{E_S\ (1-E_S)}{n}}} \\ &= \sqrt{\frac{n\ S_E\ (1-S_E)}{m\ E_S\ (1-E_S)}}.\end{aligned} \tag{5.11}$$

Para $0.3 < 1 - S_E < 0.7$ e para $0.3 < E_S < 0.7$, a equação 5.11 se aproxima do valor

$$\varrho^{'} \cong \sqrt{\frac{n}{m}}.$$

Bandas de confiança são formadas ao movimentar a curva ROC inicialmente estimada para direção noroeste e para o sudeste, ao longo de linhas com inclinação $\varrho = -\varrho^{'} = -\sqrt{n/m}$, onde n é o número total de indivíduos não doentes e m é o número total de indivíduos doentes.

Baseado em Campbell (1994), calculamos bandas globais pelo método FWB da seguinte maneira: percorremos o eixo das abscissas de $FP_i = 0$ até $FP_i = 1$ em um *grid* de valores para FP_i, de comprimento l ($l > 0$), de tal forma que $FP_i = (i-1)l$, para $i = 1, ..., 1 + \frac{1}{l}$, e encontramos o respectivo valor VP_i, para um dado valor de FP_i, usando interpolação linear, a partir da equação 5.1. Logo, para os diferentes valores de i, temos os pontos de corte (FP_i , VP_i).

A partir de uma curva ROC estimada, geramos B curvas ROC pelo método de reamostragem *bootstrap*, onde B é o número total de reamostragens realizadas.

Posteriormente, calculamos, para os diferentes valores de i, um valor denominado de VP_i'[2] e um valor denominado de VP_i''[3]. Esses valores, juntamente com os de FP_i encontrados anteriormente, para $i = 1, ..., 1 + \frac{1}{l}$, formam, respectivamente, os pontos $(FP_i\ ,\ VP_i')$ e $(FP_i\ ,\ VP_i'')$, de tal forma que a distância entre ambos os pontos com o ponto de corte $(FP_i\ ,\ VP_i)$, pertencente à curva ROC estimada, seja de

$$|(FP_i\ ,\ VP_i') - (FP_i\ ,\ VP_i)| = d = |(FP_i\ ,\ VP_i) - (FP_i\ ,\ VP_i'')|.$$

Seja $(FP_i\ ,\ VP_i)$ um ponto de corte qualquer pertencente à curva ROC estimada e $(FP_i\ ,\ VP_i')$ um ponto que pertence à fronteira superior, para $i = 1, ..., 1 + \frac{1}{l}$. Temos que a distância d entre esses dois pontos é dada por

$$(VP_i' - VP_i)^2 + (FP_i' - FP_i)^2 = d^2. \tag{5.12}$$

Seja $(FP_i\ ,\ VP_i)$ um ponto de corte qualquer pertencente à curva ROC estimada e $(FP_i\ ,\ VP_i')$ um ponto que pertence à fronteira superior, para $i = 1, ..., 1 + \frac{1}{l}$. Temos que uma reta que passa por estes pontos é dada por

$$VP_i' - VP_i = \varrho\ (FP_i' - FP_i), \tag{5.13}$$

em que ϱ é o coeficiente angular da reta (inclinação), cujo valor adotamos como sendo $\varrho = -\sqrt{n/m}$, onde m o número total de indivíduos doentes e n o número total de indivíduos não doentes.

Elevando ao quadrado ambos os lados da igualdade em 5.13, temos que

$$(VP_i' - VP_i)^2 = \varrho^2\ (FP_i' - FP_i)^2. \tag{5.14}$$

[2]: *Ponto pertencente à fronteira superior.*
[3]: *Ponto pertencente à fronteira inferior.*

Substituindo $(VP_i' - VP_i)^2$ da equação 5.14 na equação 5.12, temos que

$$\begin{aligned} \varrho^2\ (FP_i' - FP_i)^2 + (FP_i' - FP_i)^2 &= d^2 \\ (\varrho^2 + 1)(FP_i' - FP_i)^2 &= d^2 \\ (FP_i' - FP_i)^2 &= \frac{d^2}{\varrho^2 + 1}. \end{aligned} \tag{5.15}$$

Analogamente, os cálculos podem ser realizados para os valores de d e $(FP_i\ ,\ VP_i'')$, ponto este pertencente à fronteira inferior. Logo, obtemos o seguinte resultado

$$(FP_i'' - FP_i)^2 = \frac{d^2}{\varrho^2 + 1}. \tag{5.16}$$

Da equação 5.15, temos que $FP_i' - FP_i$ pode assumir valores positivos $(FP_i' - FP_i > 0)$ ou valores negativos $(FP_i' - FP_i < 0)$. Entretanto, como a inclinação é negativa $(\varrho = -\sqrt{n/m})$, temos que $FP_i' < FP_i$, para $i = 1, ..., 1 + \frac{1}{l}$.

Assim, para $FP_i' - FP_i < 0$, temos que

$$\begin{aligned} FP_i' - FP_i &= -\sqrt{\frac{d^2}{\varrho^2 + 1}} \\ FP_i' &= FP_i - \sqrt{\frac{d^2}{\varrho^2 + 1}}. \end{aligned} \tag{5.17}$$

Consequentemente, temos que

$$VP_i' = VP_i - \varrho\sqrt{\frac{d^2}{\varrho^2 + 1}}. \tag{5.18}$$

Da equação 5.16, temos que $FP_i'' - FP_i$ pode assumir valores positivos $(FP_i'' - FP_i > 0)$ ou valores negativos $(FP_i'' - FP_i < 0)$. Entretanto, como a inclinação é negativa $(\varrho = -\sqrt{n/m})$, temos que $FP_i < FP_i''$, para $i = 1, ..., 1 + \frac{1}{l}$.

Assim, para $FP_i'' - FP_i > 0$, temos que

$$FP_i'' - FP_i = \sqrt{\frac{d^2}{\varrho^2 + 1}}$$
$$FP_i'' = FP_i + \sqrt{\frac{d^2}{\varrho^2 + 1}}. \tag{5.19}$$

Consequentemente, temos que

$$VP_i'' = VP_i + \varrho\sqrt{\frac{d^2}{\varrho^2 + 1}}. \tag{5.20}$$

Após obtermos esses valores, bandas de confiança superior e inferior são encontradas. A banda superior corresponde à região que varia desde a curva ROC inicialmente estimada até a curva limite superior, dada quando unimos todos os valores $(FP_i' \ , \ VP_i')$, para $i = 1, ..., 1 + \frac{1}{l}$, ou seja, quando unimos todos os pontos

$$\left(FP_i - \sqrt{\frac{d^2}{\varrho^2 + 1}} \ , \ VP_i - \varrho\sqrt{\frac{d^2}{\varrho^2 + 1}} \right).$$

A banda inferior corresponde à região que varia desde a curva limite inferior, dada quando unimos todos os valores $(FP_i'' \ , \ VP_i'')$, para $i = 1, ..., 1 + \frac{1}{l}$, ou seja, quando unimos todos os pontos

$$\left(FP_i + \sqrt{\frac{d^2}{\varrho^2 + 1}} \ , \ VP_i + \varrho\sqrt{\frac{d^2}{\varrho^2 + 1}} \right)$$

até a curva ROC inicialmente estimada. Logo, tendo as bandas de confiança superior e inferior, temos a região de confiança para esta curva ROC.

Assim, buscamos uma distância $2d$, encontrada sobre uma certa inclinação ϱ, que contenha $(1-\alpha)$ do total das curvas ROC geradas por reamostragem *bootstrap*, onde α $(0 \leq \alpha \leq 1)$ é o nível de significância.

A Figura 5.10 ilustra essa metodologia, em que foi gerada uma região de probabilidade para a curva ROC a partir do método das bandas de largura fixa.

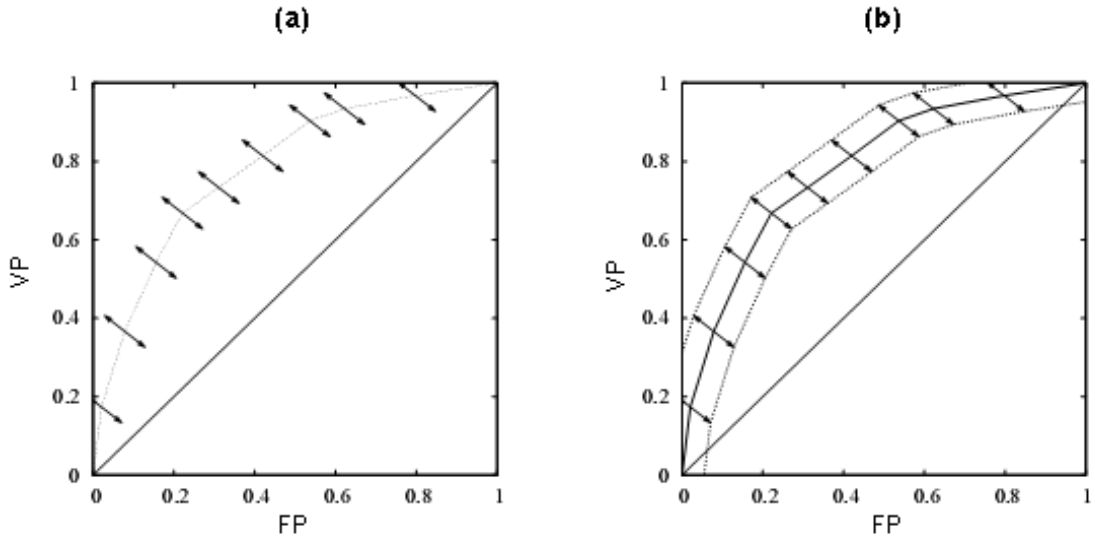

FIGURA 5.10: Ilustração do método FWB, onde (a) representa a forma de geração dos intervalos de confiança em torno dos pontos de corte (FP_i , VP_i) e (b) representa a região de incerteza relacionada com esses intervalos.

Observação:- Macskassy e Provost (2004), baseados no estudo de Campbell (1994), propõem que a inclinação seja de $\varrho = -\sqrt{VP/VN}$, onde VP é a probabilidade de um indivíduo doente ser classificado como doente e VN é a probabilidade de um indivíduo não doente ser classificado como não doente. Em nosso trabalho, usamos a inclinação $\varrho = -\sqrt{n/m}$, pois se usássemos $\varrho = -\sqrt{VP/VN}$, para cada ponto de corte estimado, teríamos diferentes valores para ϱ.

5.4 Geração de bandas de confiança

Um algoritmo geral usado em cada um desses métodos, para geração das bandas de confiança, é descrito a seguir:

1. Criar uma distribuição empírica de curvas *ROC*, caso o método necessite;

2. Escolher/gerar pontos para as regiões de confiança;

(a) Escolher o tipo do intervalo *bootstrap* (percentil, binomial, normal, studentizado) para cada um dos métodos;

(b) Percorrer os pontos de corte estimados na curva ROC dos dados originais, ou os pontos de corte da curva ROC média, quando o método necessite, a fim de que se encontre as fronteiras de confiança para cada um dos métodos;

3. Criar bandas de confiança para os intervalos encontrados no passo 2(b) para, posteriormente, encontrar a região de confiança.

Capítulo 6

Aplicação

Neste Capítulo, será apresentada uma aplicação com dados reais proposto por Zhou *et al.* (2002) sobre um exame de mamografia, com o objetivo de encontrar e comparar, especificamente para este teste diagnóstico em questão, a região de incerteza e a probabilidade de cobertura das bandas de confiança geradas pelos diferentes métodos propostos por Campbell (1994), Fawcett (2003), Macskassy *et al.* (2003) e Macskassy e Provost (2004), em torno da curva *ROC* que representa tal teste.

6.1 Descrição do conjunto de dados

Esta aplicação é apresentada inicialmente em Zhou *et al.* (2002), mas nos baseamos no exemplo encontrado no livro *Bayesian Biostatistics and Diagnostic Medicine*, de Lyle D. Broemeling, publicado em 2007.

Um estudo tem como objetivo avaliar o desempenho do exame de mamografia. As mulheres selecionadas compuseram uma amostra de tamanho 60, onde $n = 30$ nunca tiveram câncer de mama e $m = 30$ já tiveram a doença.

O padrão-ouro foi baseado no "mamograma", em que um radiologista nomeia, com uma contagem de 1 até 5, os possíveis resultados deste "mamograma", de tal forma que: 1 indica uma lesão qualquer (normal) na mama, 2 um tumor benigno, 3 uma lesão na mama que é, provavelmente, um tumor benigno,

4 indica suspeita de câncer na mama e 5 um tumor maligno. A Tabela 6.1 resume as informações fornecidas.

TABELA 6.1: Resultado do teste de mamografia

	Estado do paciente	
Resultado do teste	Câncer	Não câncer
Normal	1	9
Benigno	0	2
Provavelmente benigno	6	11
Suspeito	11	8
Maligno	12	0
Total de pacientes	30	30

6.1.1 Categorizando os resultados da mamografia

Caso o resultado do teste diagnóstico fosse binário, do tipo: positivo e negativo, nós calcularíamos a fração de verdadeiro positivo e a fração de falso positivo, encontraríamos um ponto de corte e, passando por este, estimaríamos a curva *ROC*. Como neste teste de mamografia o radiologista declarou como sendo cinco os possíveis resultados para o teste, as contagens terão que ser convertidas para um sistema binário. Logo, categorizamos os resultados da mamografia por

$$\begin{cases} \text{Teste relata normal} \rightarrow T = 1, \\ \text{Teste relata benigno} \rightarrow T = 2, \\ \text{Teste relata provavelmente benigno} \rightarrow T = 3, \\ \text{Teste relata suspeito} \rightarrow T = 4, \\ \text{Teste relata maligno} \rightarrow T = 5, \end{cases}$$

e os resultados referentes ao estado do paciente por

$$\begin{cases} \text{Paciente doente (com câncer de mama)} \rightarrow D = 1, \\ \text{Paciente não doente (sem câncer de mama)} \rightarrow D = 0. \end{cases}$$

Este conjunto de dados foi escolhido, dentre tantos outros, por ser bastante citado em publicações que envolvem a curva *ROC* para testes diagnósticos

e por apresentar características de interesse nesse estudo, tais como os problemas na estimação da região de incerteza pelos diferentes métodos propostos. Estes problemas estão relacionados com pares de variáveis categóricas que não constam nos dados levantados (não consta o par $(D\ ,\ T) = (0\ ,\ 5)$) ou que apareçam em pequena quantidade e que, durante a reamostragem, podem não constar nos dados simulados (tal como o par $(D\ ,\ T) = (1\ ,\ 1)$).

6.1.2 Pontos de corte para o teste mamográfico

Com os dados apresentados conforme a regra de classificação, para este específico teste diagnóstico, encontramos seis pontos de corte fixos, sendo dois deles os extremos da curva, que são os pontos $(0\ ,\ 0)$ e $(1\ ,\ 1)$. Logo, quatro pontos de corte "intermediários" servirão, juntamente com os dois pontos extremos, para traçar a curva *ROC*, sendo estes denominados por **A**, **B**, **C** e **D**.

Por exemplo, quando os resultados normal e benigno consideram uma resposta negativa para o teste, e os resultados provavelmente benigno, suspeito e maligno consideram uma resposta positiva para o teste, encontramos o ponto de corte **B**. Apresentamos estes resultados na Tabela 6.2.

TABELA 6.2: Ponto de corte **B**

	Câncer	**Não câncer**
Teste de mamografia (+)	29	19
Teste de mamografia (-)	1	11
	$S_E = 0.9666667$	$E_S = 0.3666667$

Seguindo o mesmo raciocínio, estimamos S_E, $1 - S_E$, E_S e $1 - E_S$ para todos os pontos de corte. Dispomos estes valores na Tabela 6.3.

TABELA 6.3: Medidas de desempenho para os pontos de corte A, B, C e D

	A	**B**	**C**	**D**
S_E	0.9666667	0.9666667	0.7666667	0.4000000
$1 - S_E$	0.0333333	0.0333333	0.2333333	0.6000000
E_S	0.3000000	0.3666667	0.7333333	1.0000000
$1 - E_S$	0.7000000	0.6333333	0.2666667	0.0000000

Assim, com as coordenadas $1 - E_S$ e S_E dos pontos de corte dispostos na Tabela 6.3, traçamos a curva ROC para o teste diagnóstico em questão.

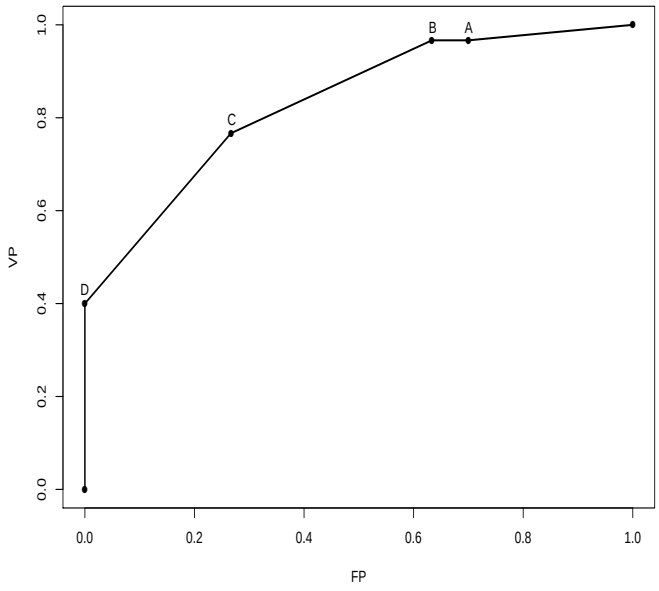

FIGURA 6.1: Curva ROC para o teste diagnóstico.

6.2 Região de incerteza

6.2.1 Metodologia

Fixamos o nível de significância em $\alpha = 0.05$, a fim de obtermos uma probabilidade de cobertura simulada próxima da cobertura nominal, cujo valor é de $(1 - \alpha) = 0.95$, para todos os métodos.

Para encontrar a região de incerteza em torno da curva ROC apresentada na Figura 6.1, reamostramos os dados originais pelo método de simulação *bootstrap* não paramétrico e realizamos: 50, 100, 500, 1000, 2500, 5000 e 10000 reamostragens. Durante cada reamostragem, fixamos uma semente no valor de 270183.

Percorremos, quando o método exige[4], ao longo das taxas de FP em intervalos regulares l de comprimento 0.01. Para melhores resultados, um *grid*

[4]*: Métodos: VA, MO, SJR e FWB.*

mais fino[5] seria o ideal, entretanto o custo computacional seria ainda mais elevado.

Usamos o *Software R 2.8.1* para a obtenção dos resultados.

6.2.2 Objetivo

Propomos uma revisão dos métodos de geração de bandas de confiança para a curva *ROC* apresentados nos trabalhos de Campbell (1994), Fawcett (2003), Macskassy *et al.* (2003) e Macskassy e Provost (2004), e elaboramos dois outros métodos de geração de bandas pontuais para a curva *ROC*. Contudo, esses autores usaram para geração das várias curvas *ROC*, que serviram para analisar o desempenho dos métodos, a validação cruzada para um grande conjunto de dados, enquanto nós usamos o processo de reamostragem *bootstrap* para um específico e pequeno conjunto de dados categóricos ordinais.

Nosso principal objetivo é verificar, para um dado nível de significância α fixo, se as bandas de confiança geradas contêm completamente as curvas *ROC* reamostradas de um mesmo modelo com uma probabilidade de cobertura de $(1-\alpha)$, ou seja, queremos verificar, para cada método, se a probabilidade de cobertura simulada esta próxima da probabilidade de cobertura esperada (nominal) e comparamos, quando for possível, esses métodos.

Observação:- Um fato que deve ser ressaltado é que **todos** os pontos de corte de **todas** as curvas *ROC* geradas pelo *bootstrap* devem estar completamente dentro da região de incerteza encontrada, para conseguirmos obter o valor da probabilidade de cobertura simulada.

6.2.3 Resultados

Bandas de confiança para o método das médias verticais (VA)

No método VA geramos intervalos unidimensionais para cada ponto de corte estimado na curva *ROC* média através de intervalos *bootstrap* percentis, binomiais e normais.

[5]*: A distância entre dois pontos sucessivos é muito pequena.*

Os intervalos *bootstrap* percentis foram gerados de tal forma que os $\frac{\alpha}{2}$-ésimo e $(1-\frac{\alpha}{2})$-ésimo quantis da distribuição empírica das curvas ROC reamostradas, organizados em forma crescente, fossem estimados.

Seja Y uma variável aleatória que representa o número de indivíduos, dentre todos os doentes, que apresenta resultado positivo no teste diagnóstico. Vamos supor que $Y \sim Bin\ (m\ ,\ VP)$, onde m é o número total de indivíduos doentes e VP é a probabilidade de um indivíduo doente ser classificado como doente. Assim, em cada ponto de corte estimado na curva ROC média, encontramos intervalos *bootstrap* binomiais para os valores de VP.

Embora citamos, ao longo do texto, que uma das formas de encontrar bandas pontuais pelo método VA usa a distribuição normal, na realidade estamos estimando intervalos studentizados, pois o tamanho amostral é reduzido (m = 30 indivíduos doentes). Assim, em cada ponto de corte estimado na curva ROC média, encontramos intervalos *bootstrap* studentizados para os valores de VP a partir de uma distribuição t-Student com 29 graus de liberdade.

Nas Figuras 6.2, 6.3, 6.4 e 6.5 dispomos as curvas ROC geradas por reamostragem *bootstrap* para o método VA, bem como suas bandas de confiança estimadas pelas distribuições empírica, binomial e normal.

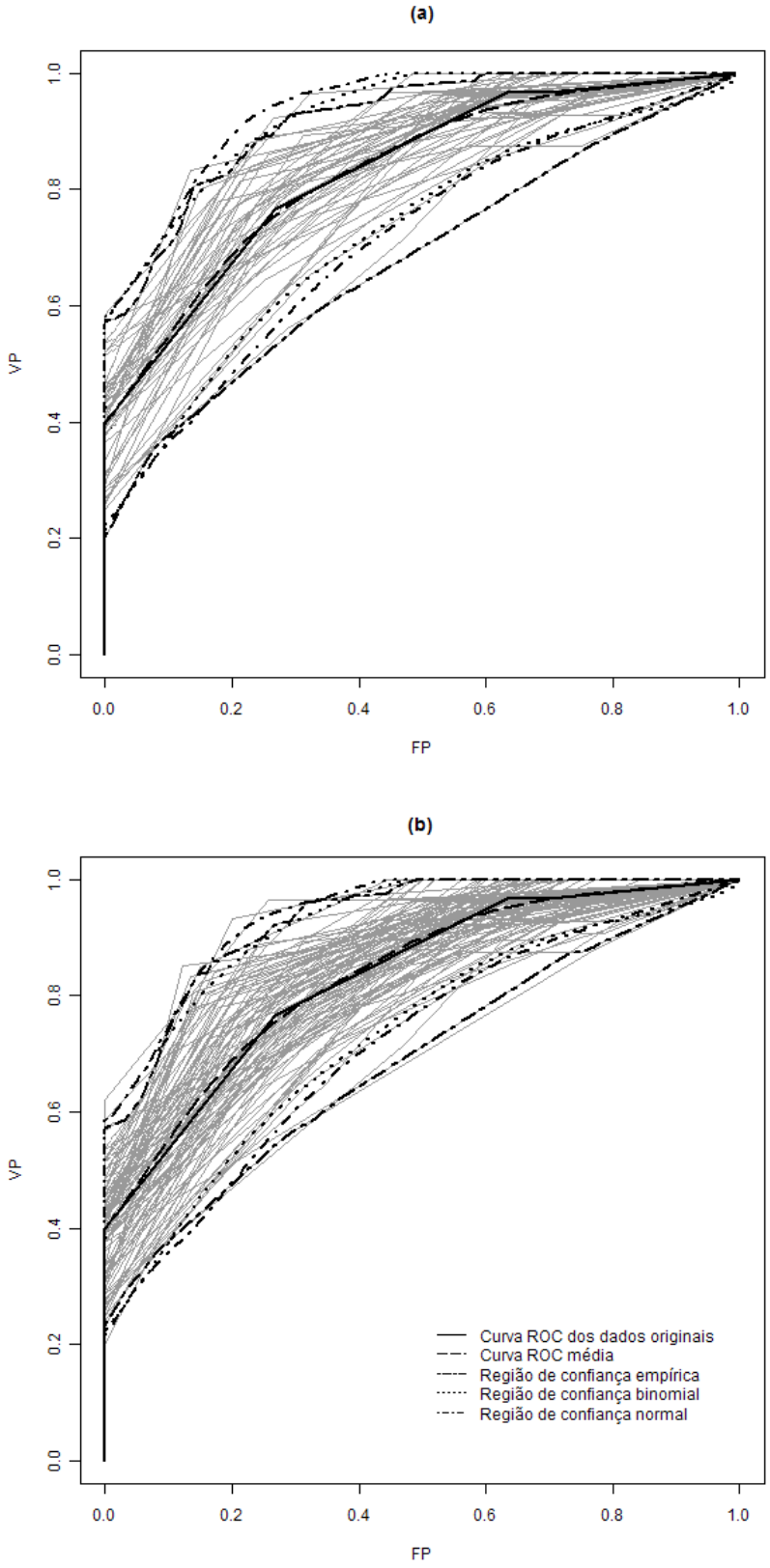

FIGURA 6.2: Bandas de confiança estimadas pelo método VA para: (a) 50 reamostragens e (b) 100 reamostragens.

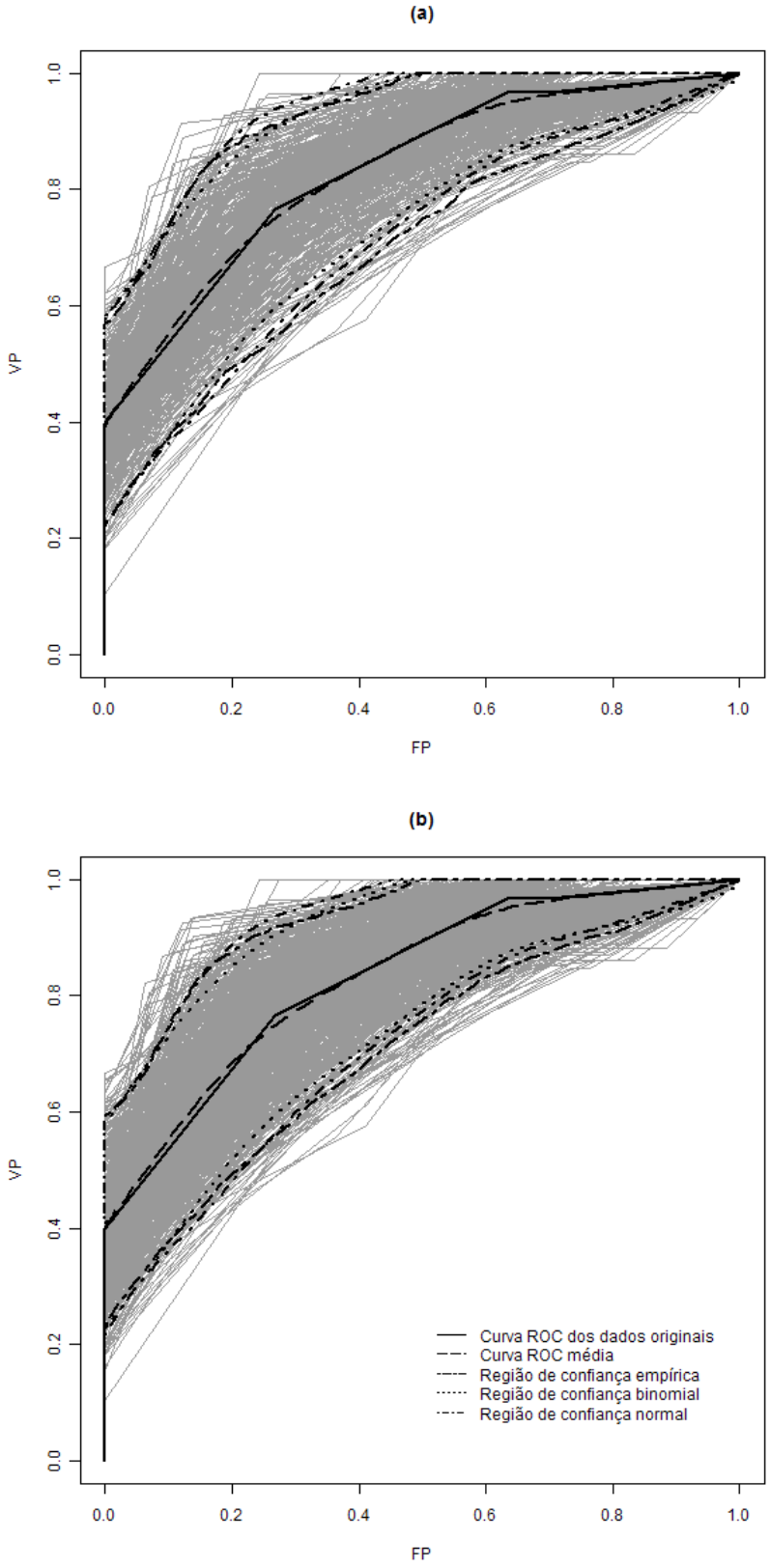

FIGURA 6.3: Bandas de confiança estimadas pelo método VA para: (a) 500 reamostragens e (b) 1000 reamostragens.

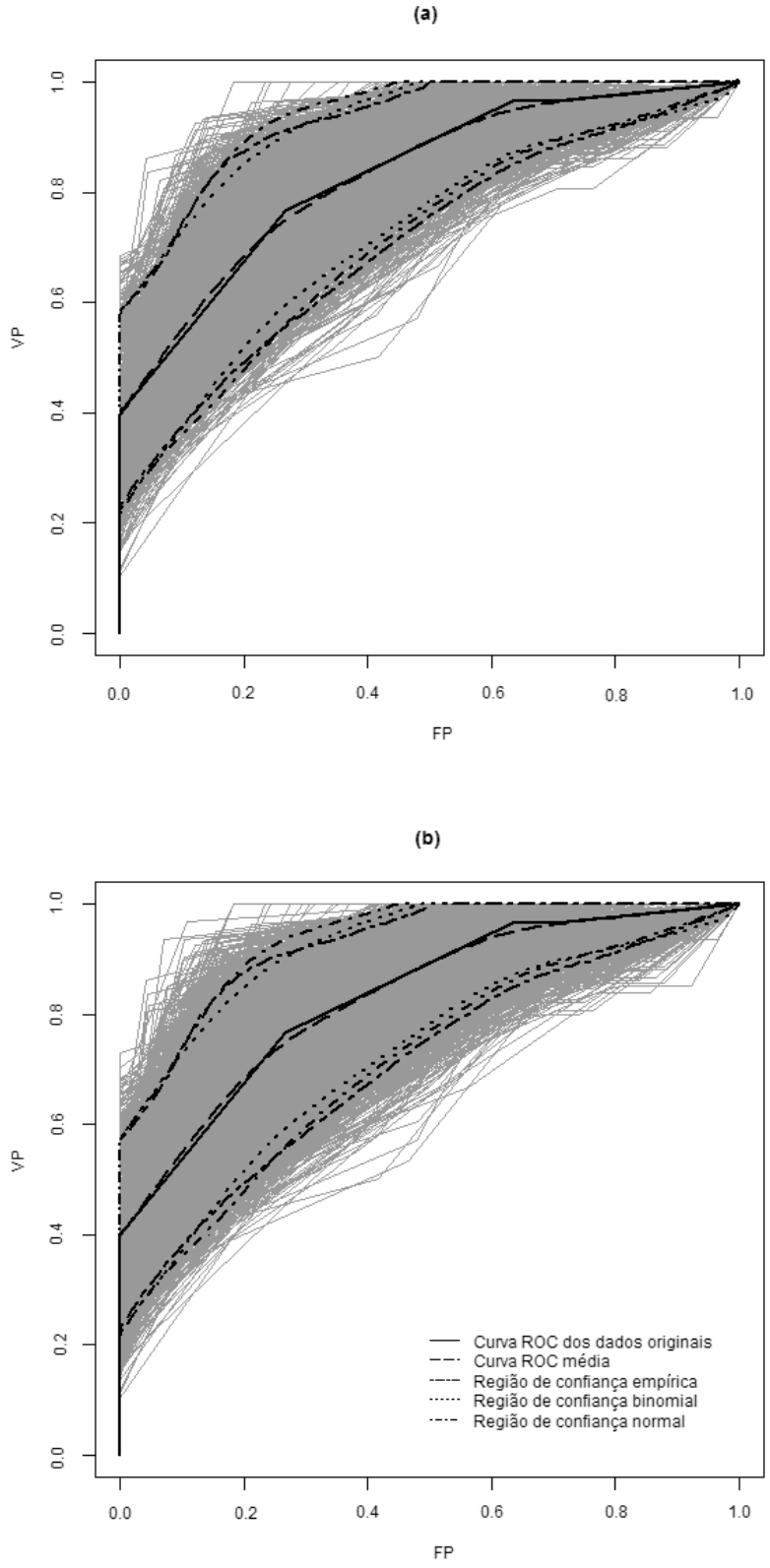

FIGURA 6.4: Bandas de confiança estimadas pelo método VA para: (a) 2500 reamostragens e (b) 5000 reamostragens.

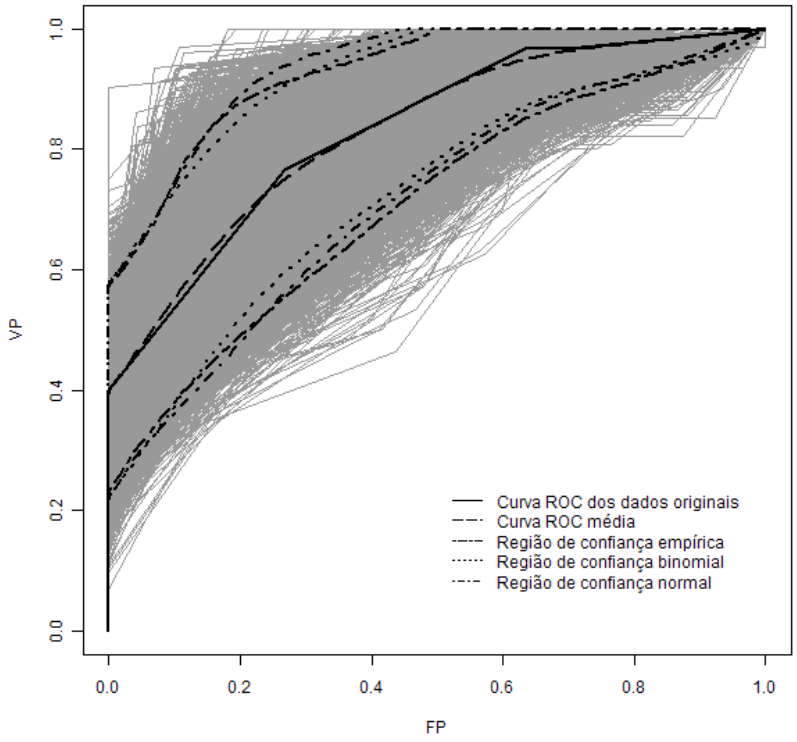

FIGURA 6.5: Bandas de confiança estimadas pelo método VA para 10000 reamostragens.

Analisando as Figuras 6.2, 6.3, 6.4 e 6.5 podemos observar que os intervalos de confiança verticais encontrados no método VA são mais largos para menores valores de FP.

Após realizadas todas as B reamostrangens *bootstrap* e obtidas as curvas ROC para cada reamostragem, encontramos a proporção das curvas ROC que estão completamente dentro das bandas de confiança estimadas pelo método VA, para cada uma das distribuições assumidas. Na Tabela 6.4 dispomos os valores da probabilidade de cobertura simulada por esse método.

TABELA 6.4: Probabilidade de cobertura das bandas de confiança VA

	Número de reamostragens *bootstrap*						
Probabilidade de cobertura	**50**	**100**	**500**	**1000**	**2500**	**5000**	**10000**
Empírica	0.9	0.89	0.832	0.824	0.8108	0.8062	0.8064
Binomial	0.72	0.77	0.764	0.752	0.7488	0.7508	0.7508
Normal	0.84	0.84	0.83	0.821	0.8256	0.8256	0.8234

Na Figura 6.6, com os valores dispostos na Tabela 6.4, construímos o gráfico das probabilidades de cobertura simulada *versus* a quantidade de reamostragens *bootstrap*.

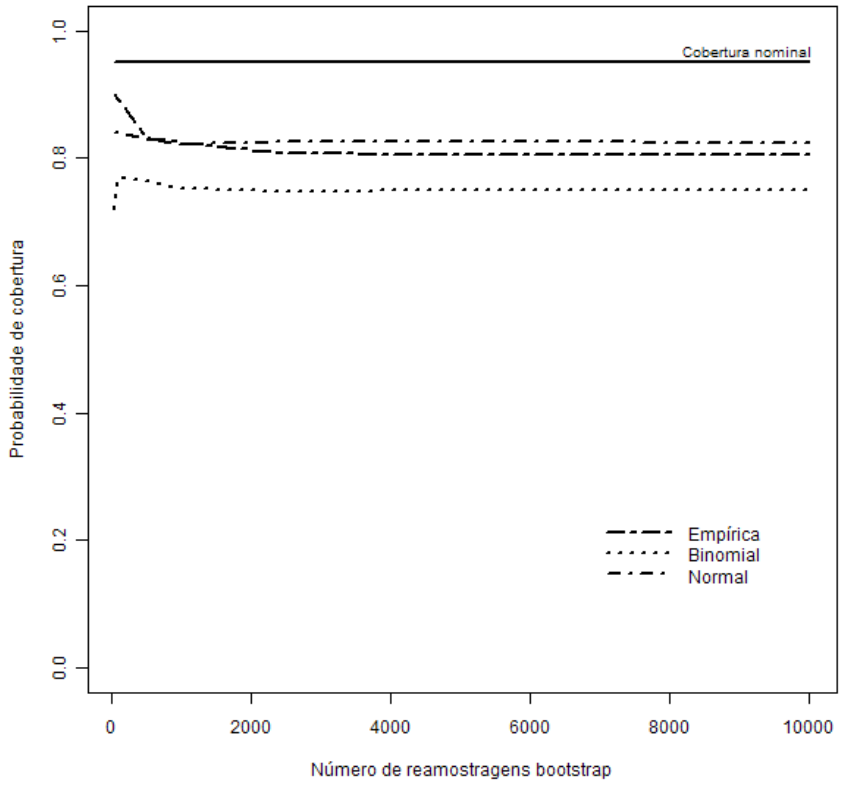

FIGURA 6.6: Probabilidades de cobertura para o método VA.

Analisando a Tabela 6.4 e a Figura 6.6 notamos que o valor da probabilidade de cobertura simulada é inferior ao valor da probabilidade nominal (0.95), para as três distribuições assumidas na geração de intervalos pelo método VA, independentemente da quantidade de reamostragens *bootstrap* efetuada. Assim, a probabilidade de cobertura simulada subestima o valor da probabilidade de cobertura.

Para bandas geradas pela distribuição empírica, quando efetuado um número reduzido de reamostragens, a cobertura é em torno de 90% das curvas *ROC* geradas. Entretanto, conforme o número de reamostragens aumenta, a cobertura de tais bandas diminui e se estabiliza em torno de 80% do total das curvas geradas.

Também notamos que para a distribuição binomial, a probabilidade de cobertura simulada é a menor de todas as coberturas estimadas, quando comparada com as outras duas distribuições assumidas, cobrindo aproximadamente 75% das curvas ROC geradas.

A probabilidade de cobertura das bandas geradas através da distribuição normal, independentemente do número de reamostragens, é de aproximadamente 82% do total de curvas ROC geradas. Este é o método que mais cobre as curvas ROC reamostradas e também é o mais estável, pois o valor da cobertura das bandas normais pouco oscila conforme varia o número de reamostragens *bootstrap*.

Bandas de confiança para o método das médias otimizadas (MO)

No método MO geramos intervalos unidimensionais para cada ponto de corte estimado na curva ROC através de intervalos *bootstrap* percentis, binomiais e normais. O método MO gera, inicialmente, intervalos verticais para as taxas de VP e, posteriormente, gera intervalos horizontais para as taxas de FP, independentemente das distribuições assumidas.

Os intervalos *bootstrap* percentis foram gerados de tal forma que os $\frac{\alpha}{2}$-ésimo e $(1-\frac{\alpha}{2})$-ésimo quantis da distribuição empírica das curvas ROC reamostradas, organizados em forma crescente, fossem estimados.

Seja Y uma variável aleatória que representa o número de indivíduos, dentre todos os doentes, que apresenta resultado positivo no teste diagnóstico. Vamos supor que $Y \sim Bin\ (m\ ,\ VP)$, onde m é o número total de indivíduos doentes e VP é a probabilidade de um indivíduo doente ser classificado como doente. Assim, em cada ponto de corte estimado na curva ROC inicial, encontramos intervalos *bootstrap* binomiais para os valores de VP. Seja X uma variável aleatória que representa o número de indivíduos, dentre todos os não doentes, que apresenta resultado positivo no teste diagnóstico. Vamos supor que $X \sim Bin\ (n\ ,\ FP)$, onde n é o número total de indivíduos não doentes e FP é a probabilidade de um indivíduo não doente ser classificado como doente. Assim, em cada ponto de corte estimado na curva ROC inicial, encontramos intervalos *bootstrap* binomiais para os valores de FP.

Embora citamos, ao longo do texto, que uma das formas de encontrar bandas pontuais pelo método MO usa a distribuição normal, na realidade estamos estimando intervalos studentizados, pois o tamanho amostral é reduzido (m = 30 indivíduos doentes e n = 30 indivíduos não doentes). Assim, em cada ponto de corte estimado na curva *ROC*, inicialmente encontramos intervalos *bootstrap* studentizados para os valores de VP a partir de uma distribuição t-Student com 29 graus de liberdade, e posteriormente encontramos intervalos *bootstrap* studentizados para os valores de FP a partir de uma distribuição t-Student com 29 graus de liberdade.

Nas Figuras 6.7, 6.8, 6.9 e 6.10 dispomos as curvas *ROC* geradas por reamostragem *bootstrap* para o método MO, bem como suas bandas de confiança estimadas pelas distribuições empírica, binomial e normal.

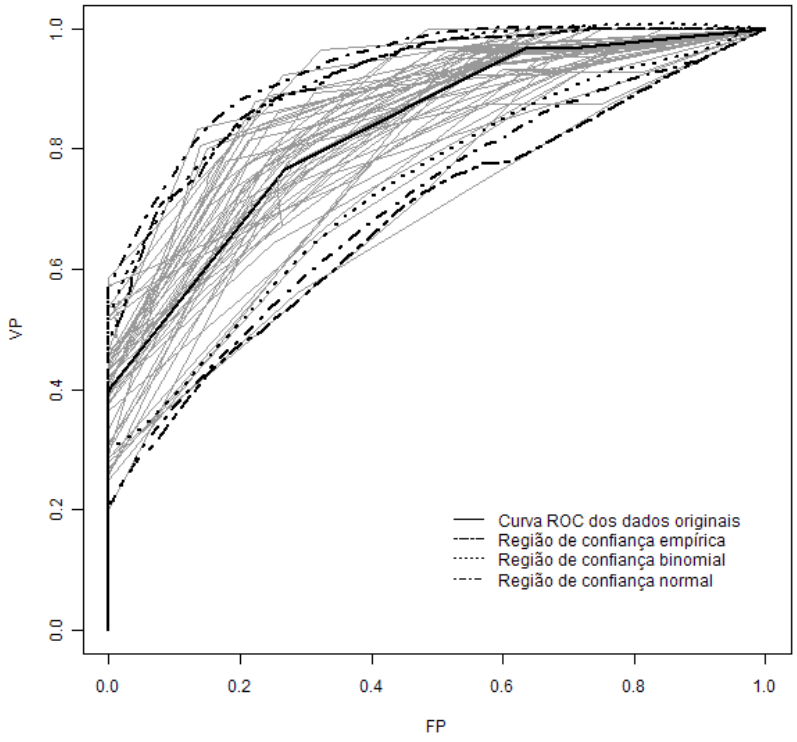

FIGURA 6.7: Bandas de confiança estimadas pelo método MO para 50 reamostragens.

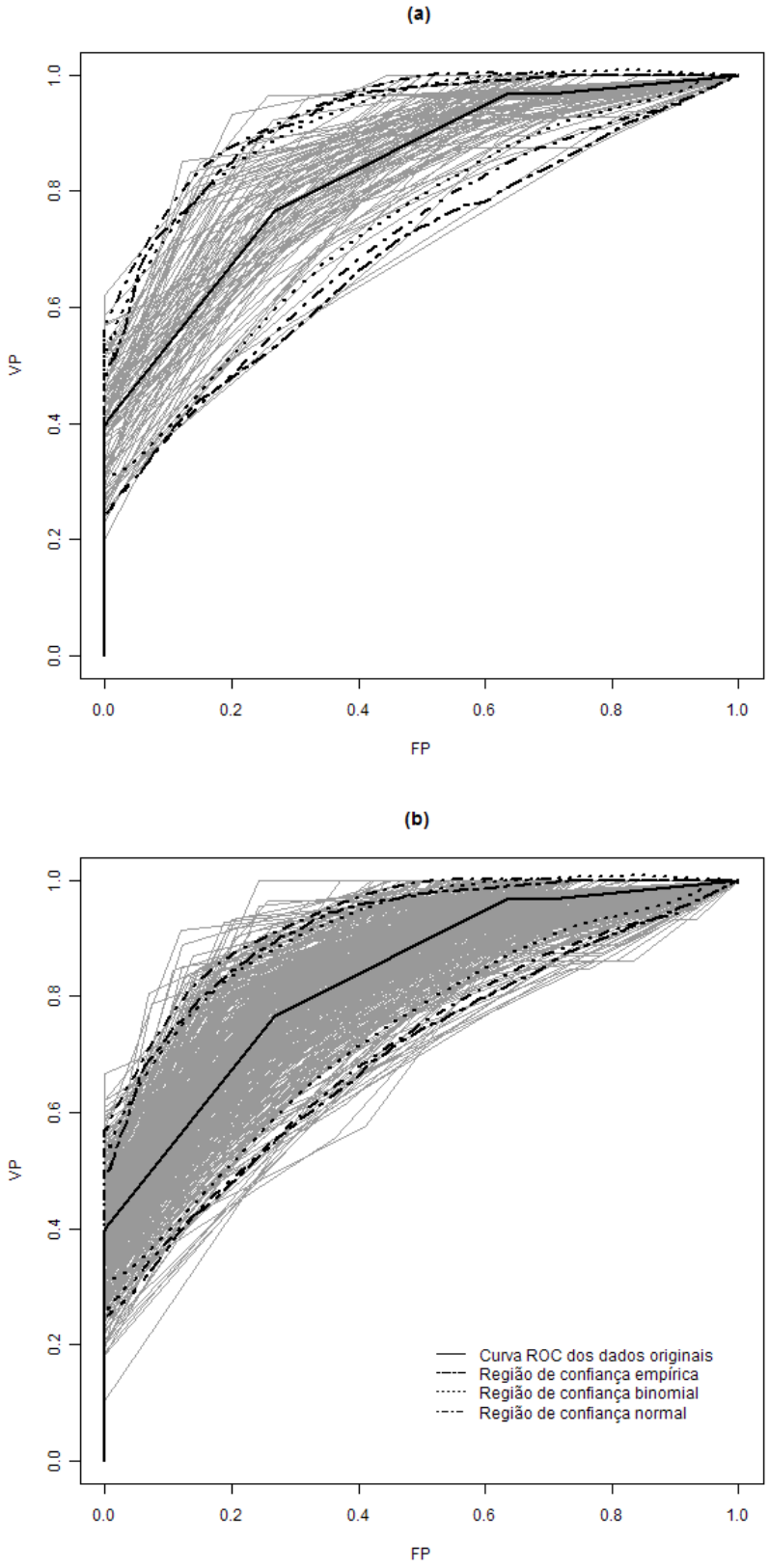

FIGURA 6.8: Bandas de confiança estimadas pelo método MO para: (a) 100 reamostragens e (b) 500 reamostragens.

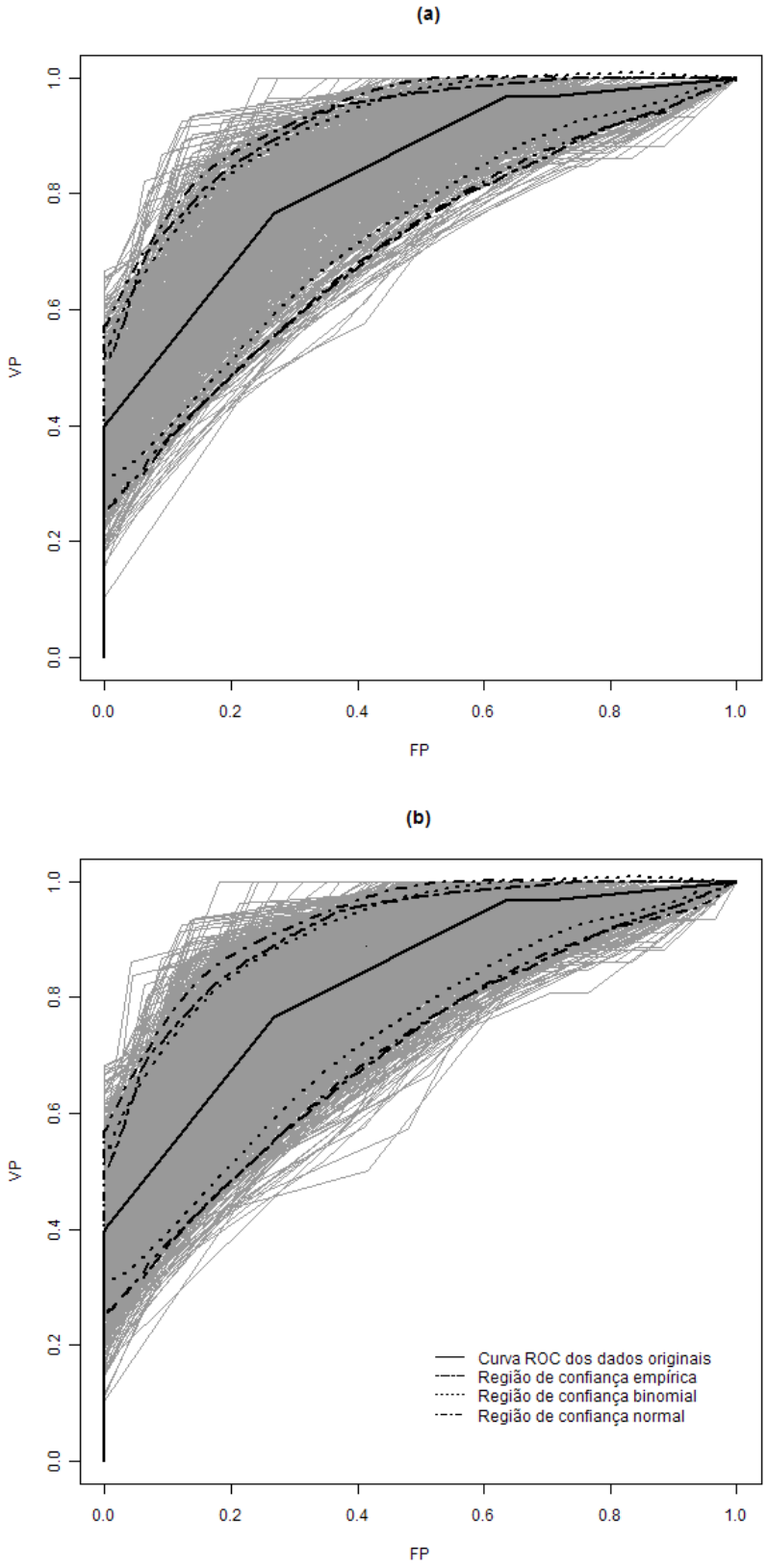

FIGURA 6.9: Bandas de confiança estimadas pelo método MO para: (a) 1000 reamostragens e (b) 2500 reamostragens.

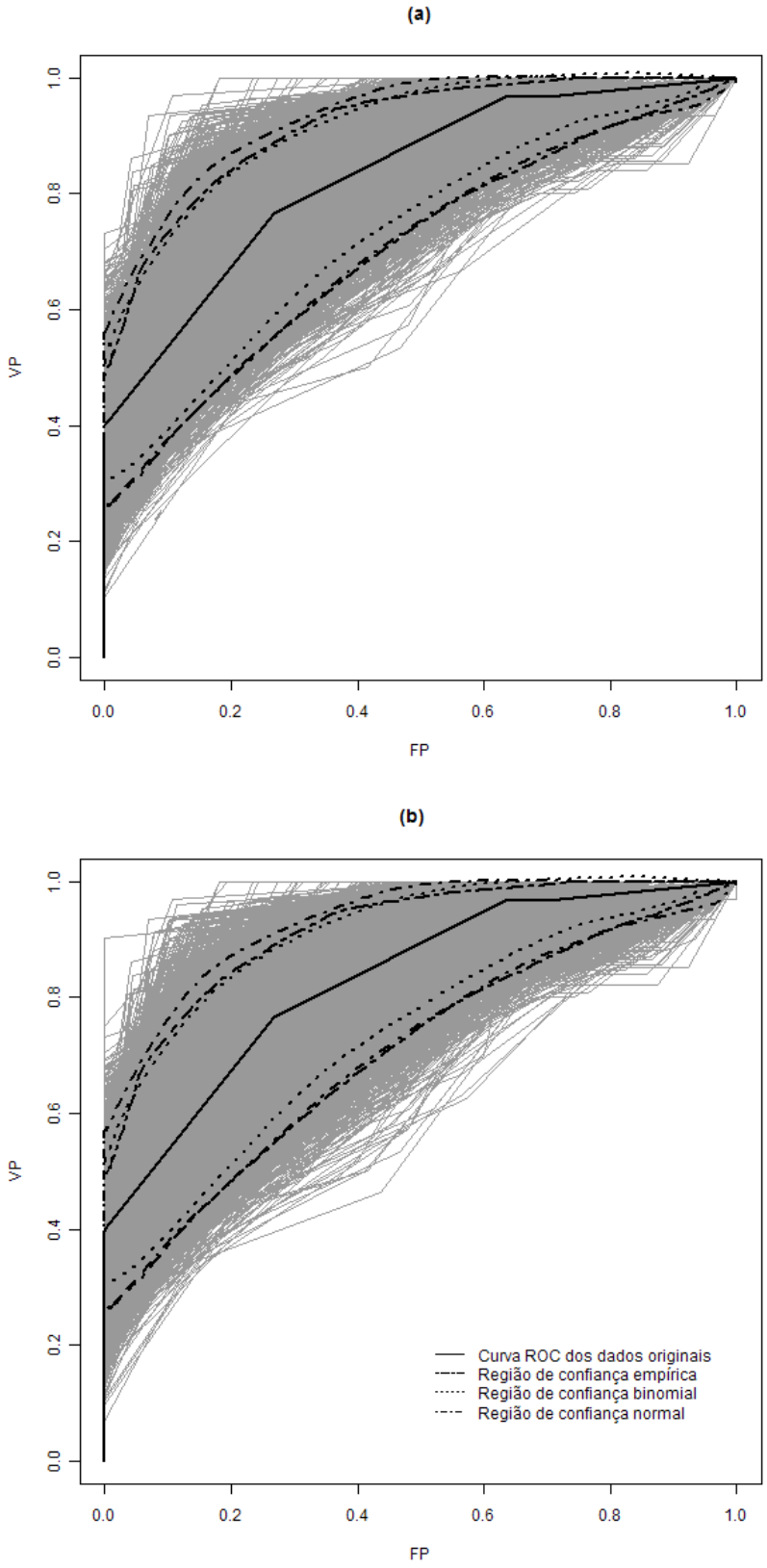

FIGURA 6.10: Bandas de confiança estimadas pelo método MO para: (a) 5000 reamostragens e (b) 10000 reamostragens.

Após realizadas todas as B reamostrangens *bootstrap* e obtidas as curvas ROC para cada reamostragem, encontramos a proporção das curvas ROC que estão completamente dentro das bandas de confiança estimadas pelo método MO, para cada uma das distribuições assumidas. Na Tabela 6.5 dispomos os valores da probabilidade de cobertura simulada por esse método.

TABELA 6.5: Probabilidade de cobertura das bandas de confiança MO

	Número de reamostragens *bootstrap*						
Probabilidade de cobertura	**50**	**100**	**500**	**1000**	**2500**	**5000**	**10000**
Empírica	0.56	0.51	0.548	0.547	0.5368	0.536	0.5341
Binomial	0.64	0.63	0.612	0.605	0.5992	0.5936	0.5993
Normal	0.84	0.83	0.844	0.836	0.8268	0.8328	0.8176

Na Figura 6.11, com os valores dispostos na Tabela 6.5, construímos o gráfico das probabilidades de cobertura simulada *versus* a quantidade de reamostragens *bootstrap*.

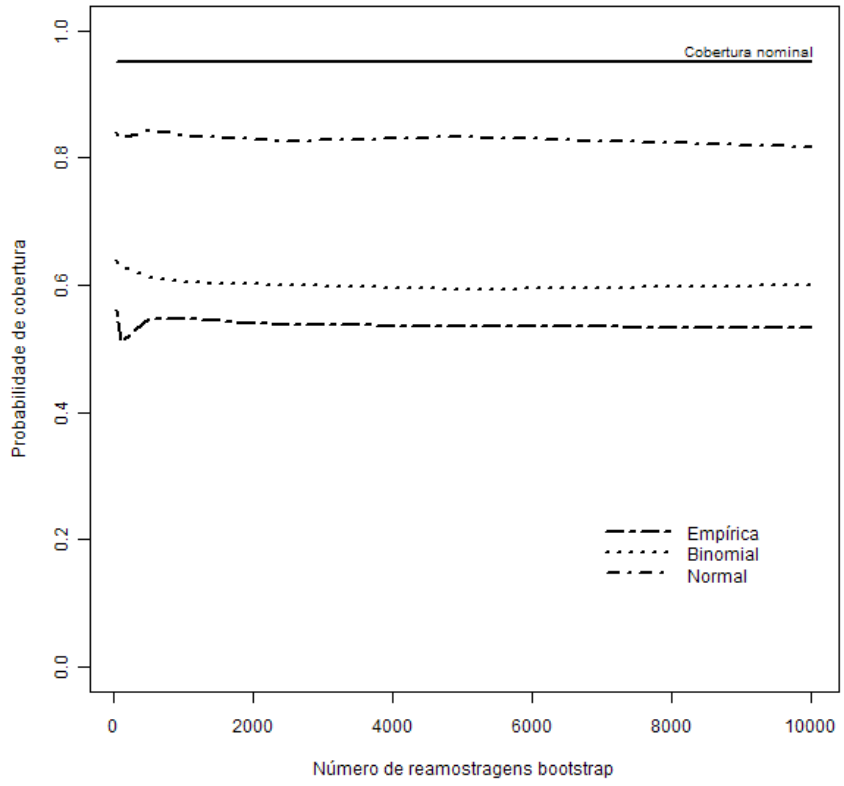

FIGURA 6.11: Probabilidades de cobertura para o método MO.

Analisando a Tabela 6.5 e a Figura 6.11 notamos que o valor da probabilidade de cobertura simulada é inferior ao valor da probabilidade nominal (0.95), para as três distribuições assumidas na geração de intervalos pelo método MO, independentemente da quantidade de reamostragens *bootstrap* efetuada. Assim, a probabilidade de cobertura simulada subestima o valor da probabilidade de cobertura.

Notamos que para a distribuição empírica, a probabilidade de cobertura simulada é a menor de todas as coberturas estimadas, quando comparado com as outras duas distribuições assumidas, cobrindo aproximadamente 53% das curvas ROC geradas.

Para bandas geradas pela distribuição binomial, quando efetuado um número reduzido de reamostragens, a cobertura é em torno de 63% das curvas ROC geradas. Entretanto, conforme o número de reamostragens aumenta, a cobertura de tais bandas diminui e se estabiliza em torno de 60% do total das curvas geradas.

A probabilidade de cobertura das bandas geradas através da distribuição normal, independentemente do número de reamostragens, é de aproximadamente 83% do total de curvas ROC geradas. Este é o método que mais cobre as curvas ROC reamostradas.

Bandas de confiança para o método das médias limiares (TA)

No método TA geramos intervalos unidimensionais para cada ponto de corte estimado na curva ROC média através de intervalos *bootstrap* percentis, binomiais e normais.

Os intervalos *bootstrap* percentis foram gerados de tal forma que os $\frac{\alpha}{2}$-ésimo e $(1-\frac{\alpha}{2})$-ésimo quantis da distribuição empírica das curvas ROC reamostradas, organizados em forma crescente, fossem estimados.

Seja Y uma variável aleatória que representa o número de indivíduos, dentre todos os doentes, que apresenta resultado positivo no teste diagnóstico. Vamos supor que $Y \sim Bin\ (m\ ,\ VP)$, onde m é o número total de indivíduos

doentes e VP é a probabilidade de um indivíduo doente ser classificado como doente. Assim, em cada ponto de corte estimado na curva ROC média, encontramos intervalos *bootstrap* binomiais para os valores de VP.

Embora citamos, ao longo do texto, que uma das formas de encontrar bandas pontuais pelo método TA usa a distribuição normal, na realidade estamos estimando intervalos studentizados, pois o tamanho amostral é reduzido (m = 30 indivíduos doentes). Assim, em cada ponto de corte estimado na curva ROC média, encontramos intervalos *bootstrap* studentizados para os valores de VP a partir de uma distribuição t-Student com 29 graus de liberdade.

Nas Figuras 6.12, 6.13, 6.14 e 6.15 dispomos as curvas ROC geradas por reamostragem *bootstrap* para o método TA, bem como suas bandas de confiança estimadas pelas distribuições empírica, binomial e normal.

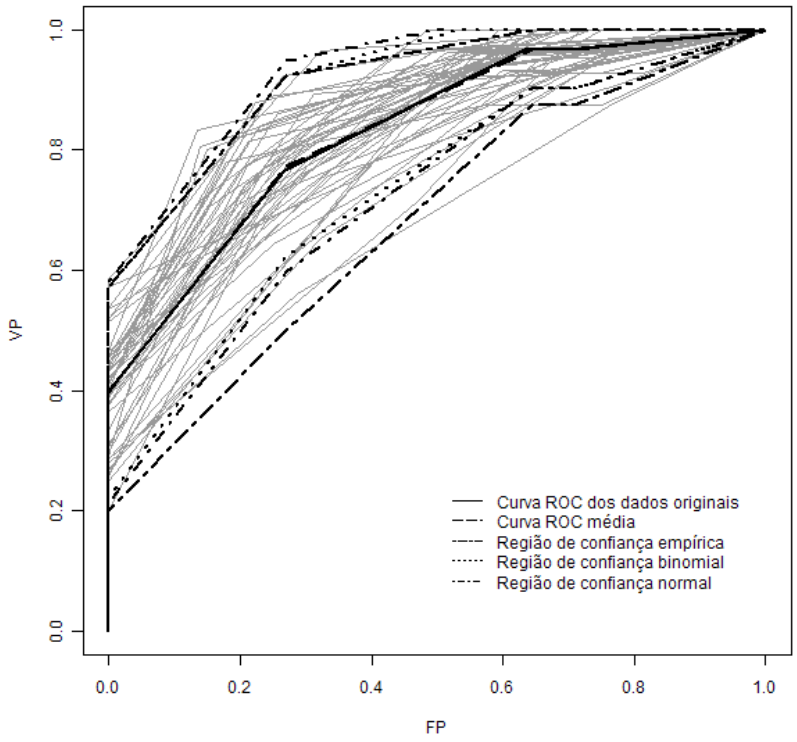

FIGURA 6.12: Bandas de confiança estimadas pelo método TA para 50 reamostragens.

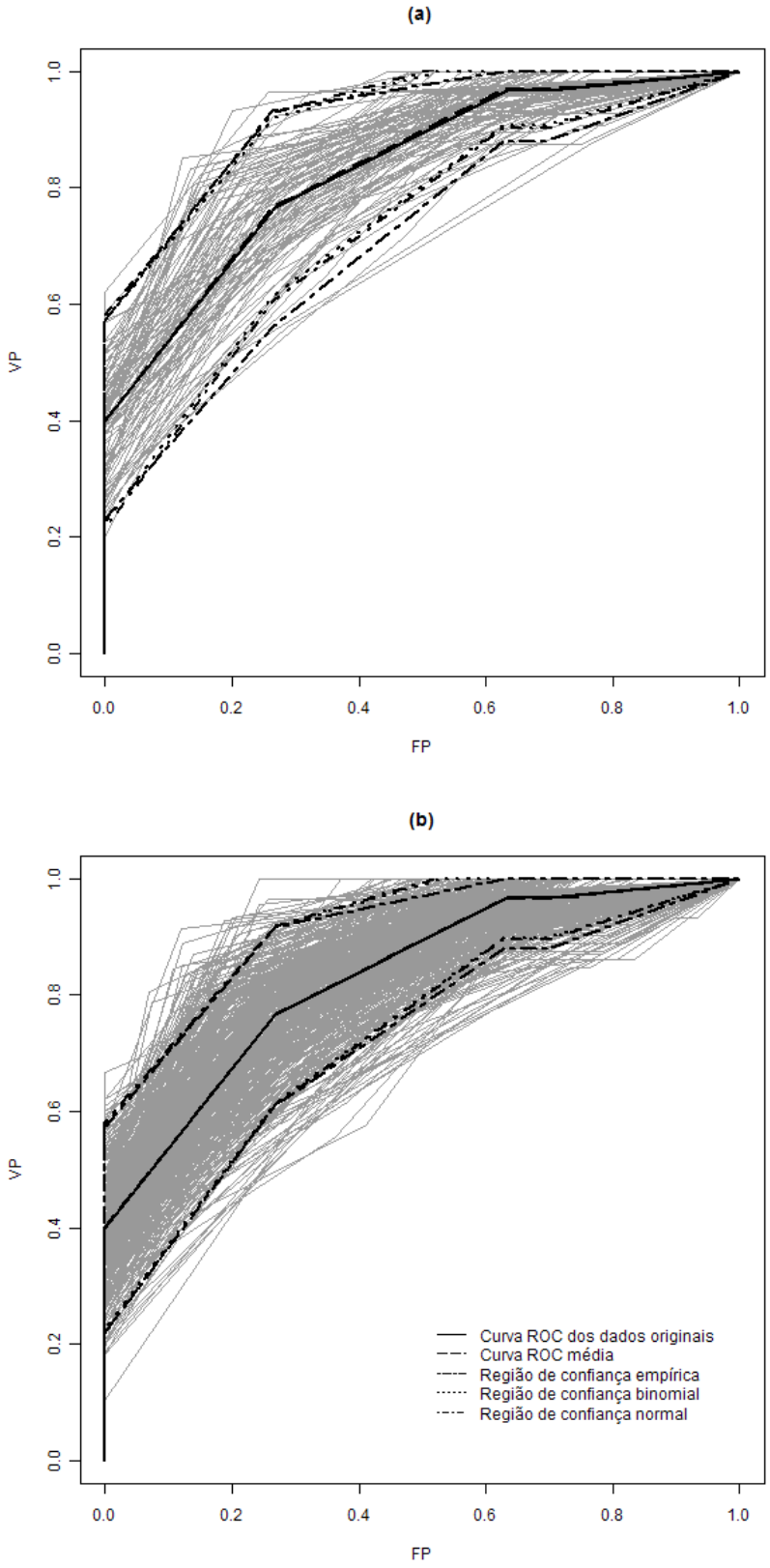

FIGURA 6.13: Bandas de confiança estimadas pelo método TA para: (a) 100 reamostragens e (b) 500 reamostragens.

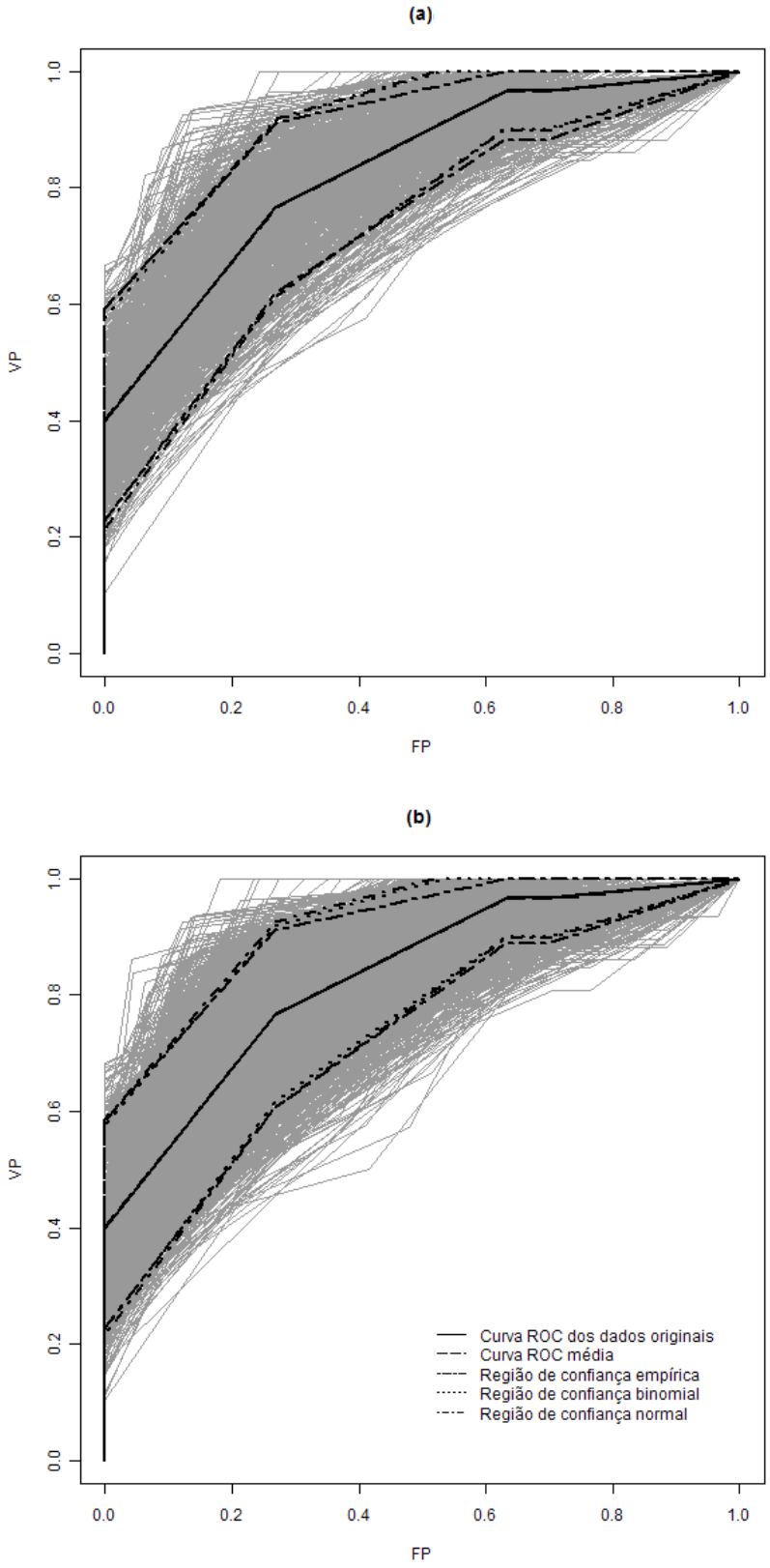

FIGURA 6.14: Bandas de confiança estimadas pelo método TA para: (a) 1000 reamostragens e (b) 2500 reamostragens.

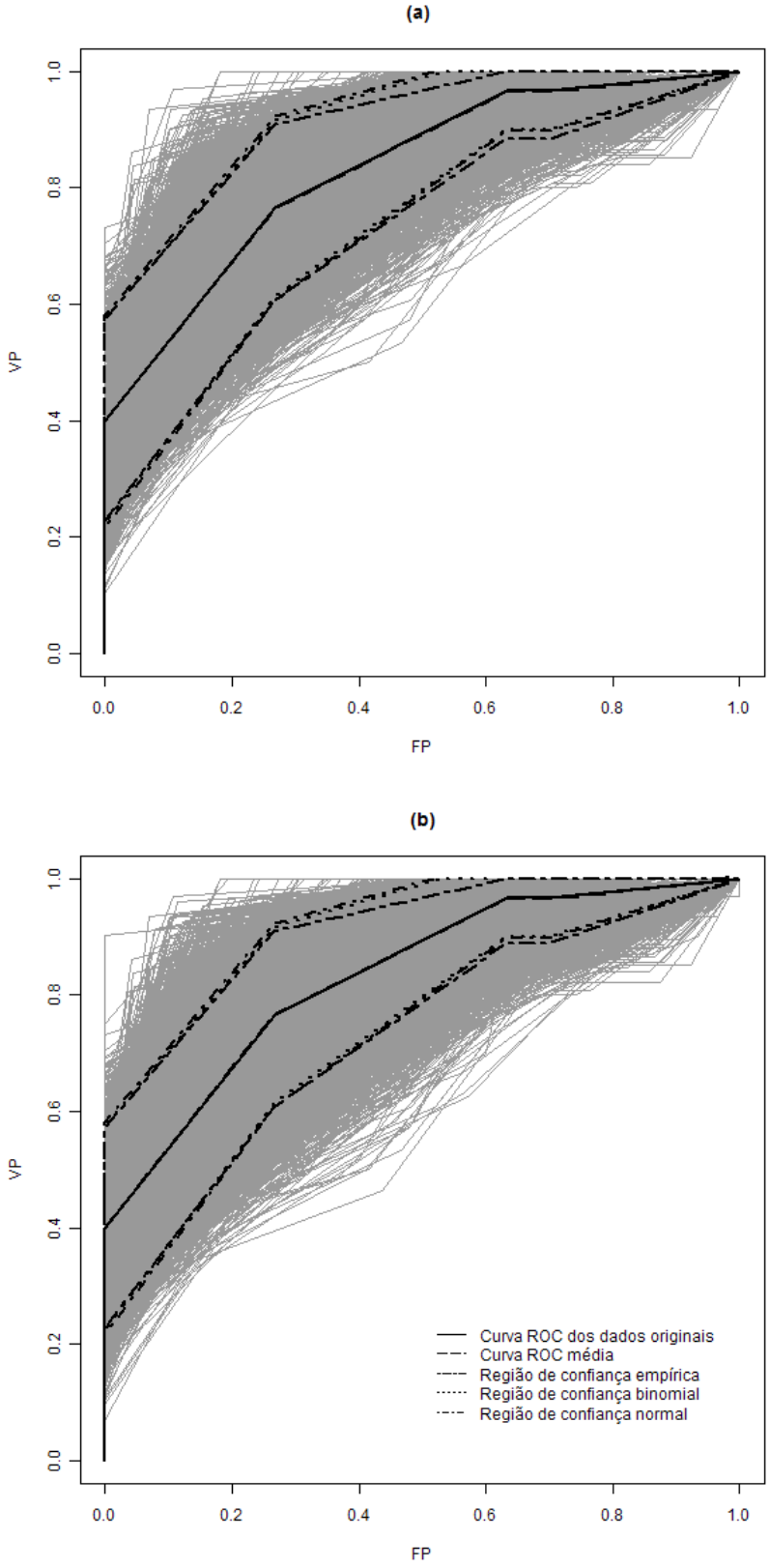

FIGURA 6.15: Bandas de confiança estimadas pelo método TA para: (a) 5000 reamostragens e (b) 10000 reamostragens.

Após realizadas todas as B reamostrangens *bootstrap* e obtidas as curvas ROC para cada reamostragem, encontramos a proporção das curvas ROC que estão completamente dentro das bandas de confiança estimadas pelo método TA, para cada uma das distribuições assumidas. Na Tabela 6.6 dispomos os valores da probabilidade de cobertura simulada por esse método.

TABELA 6.6: Probabilidade de cobertura das bandas de confiança TA

	Número de reamostragens *bootstrap*						
Probabilidade de cobertura	**50**	**100**	**500**	**1000**	**2500**	**5000**	**10000**
Empírica	0.68	0.66	0.62	0.598	0.5852	0.5936	0.5868
Binomial	0.68	0.67	0.674	0.672	0.668	0.6772	0.6756
Normal	0.78	0.72	0.692	0.701	0.7124	0.7102	0.7112

Na Figura 6.16, com os valores dispostos na Tabela 6.6, construímos o gráfico das probabilidades de cobertura simulada *versus* a quantidade de reamostragens *bootstrap*.

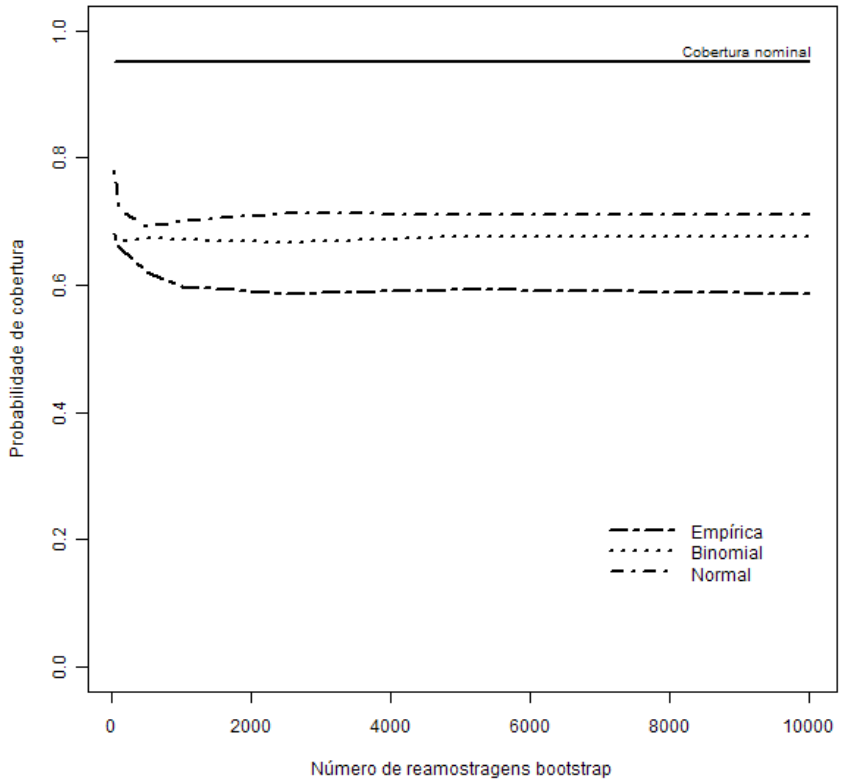

FIGURA 6.16: Probabilidades de cobertura para o método TA.

Analisando a Tabela 6.6 e a Figura 6.16 notamos que o valor da probabilidade de cobertura simulada é inferior ao valor da probabilidade nominal (0.95), para as três distribuições assumidas na geração de intervalos pelo método TA, independentemente da quantidade de reamostragens *bootstrap* efetuada. Assim, a probabilidade de cobertura simulada subestima o valor da probabilidade de cobertura.

Para bandas geradas pela distribuição empírica, quando efetuado um número reduzido de reamostragens, a cobertura é em torno de 66% das curvas *ROC* geradas. Entretanto, conforme o número de reamostragens aumenta, a cobertura de tais bandas diminui e se estabiliza em torno de 59% do total das curvas geradas. Também notamos que para a distribuição empírica, a probabilidade de cobertura simulada é a menor de todas as coberturas estimadas, quando comparada com as outras duas distribuições adotadas por este método.

Para a distribuição binomial, a probabilidade de cobertura simulada encontrada é a mais estável em relação às coberturas obtidas pelas outras distribuições, pois seu valor pouco oscila conforme varia o número de reamostragens *bootstrap*, sendo praticamente constante em 68% do total das curvas *ROC* reamostradas.

A probabilidade de cobertura das bandas geradas através da distribuição normal, quando efetuado um número reduzido de reamostragens, é em torno de 78% das curvas *ROC* geradas. Entretanto, conforme o número de reamostragens aumenta, a cobertura de tais bandas diminui e se estabiliza em torno de 71% do total das curvas *ROC* geradas. Este é o método que mais cobre as curvas *ROC* reamostradas pelo *bootstrap*.

Bandas de confiança para o método das médias limiares otimizadas (MLO)

No método MLO geramos intervalos unidimensionais nas proximidades dos pontos de corte obtidos da curva *ROC* inicialmente estimada através de intervalos *bootstrap* percentis, binomiais e normais. O método MLO gera intervalos verticais para as taxas de VP e gera intervalos horizontais para as taxas de FP,

independentemente das distribuições assumidas.

Os intervalos *bootstrap* percentis foram gerados de tal forma que os $\frac{\alpha}{2}$-ésimo e $(1-\frac{\alpha}{2})$-ésimo quantis da distribuição empírica das curvas ROC reamostradas, organizados em forma crescente, fossem estimados.

Seja Y uma variável aleatória que representa o número de indivíduos, dentre todos os doentes, que apresenta resultado positivo no teste diagnóstico. Vamos supor que $Y \sim Bin\ (m\ ,\ VP)$, onde m é o número total de indivíduos doentes e VP é a probabilidade de um indivíduo doente ser classificado como doente. Assim, próximo ao ponto de corte estimado na curva ROC inicial, encontramos intervalos *bootstrap* binomiais para os valores de VP. Seja X uma variável aleatória que representa o número de indivíduos, dentre todos os não doentes, que apresenta resultado positivo no teste diagnóstico. Vamos supor que $X \sim Bin\ (n\ ,\ FP)$, onde n é o número total de indivíduos não doentes e FP é a probabilidade de um indivíduo não doente ser classificado como doente. Assim, próximo ao ponto de corte estimado na curva ROC inicial, encontramos intervalos *bootstrap* binomiais para os valores de FP.

Embora citamos, ao longo do texto, que uma das formas de encontrar bandas pontuais pelo método MLO usa a distribuição normal, na realidade estamos estimando intervalos studentizados, pois o tamanho amostral é reduzido (m = 30 indivíduos doentes e n = 30 indivíduos não doentes). Assim, próximo de cada ponto de corte estimado na curva ROC, encontramos intervalos *bootstrap* studentizados para os valores de VP a partir de uma distribuição t-Student com 29 graus de liberdade e encontramos intervalos *bootstrap* studentizados para os valores de FP a partir de uma distribuição t-Student com 29 graus de liberdade.

Nas Figuras 6.17, 6.18, 6.19 e 6.20 dispomos as curvas ROC geradas por reamostragem *bootstrap* para o método MLO, bem como suas bandas de confiança estimadas pelas distribuições empírica, binomial e normal.

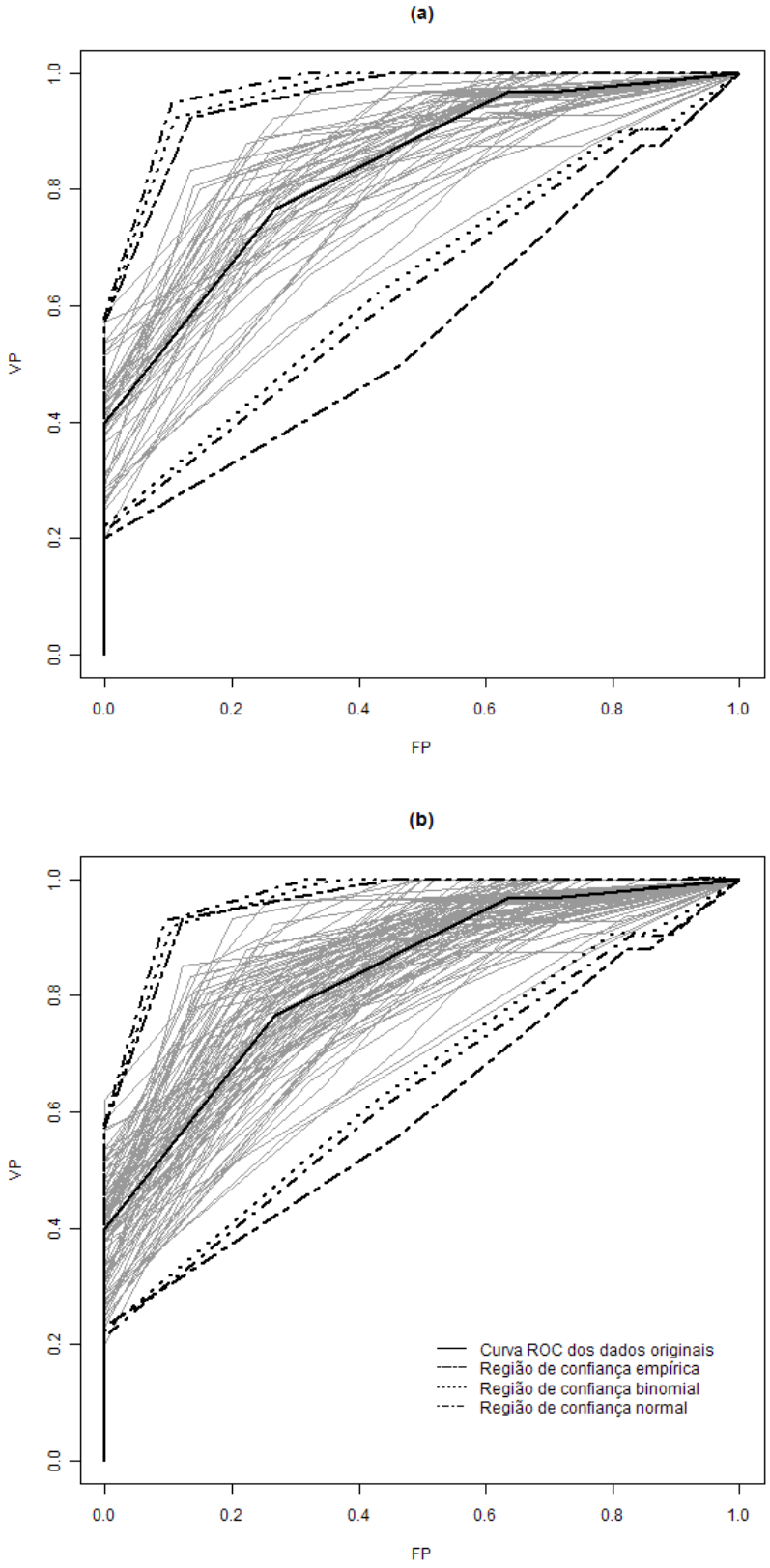

FIGURA 6.17: Bandas de confiança estimadas pelo método MLO para: (a) 50 reamostragens e (b) 100 reamostragens.

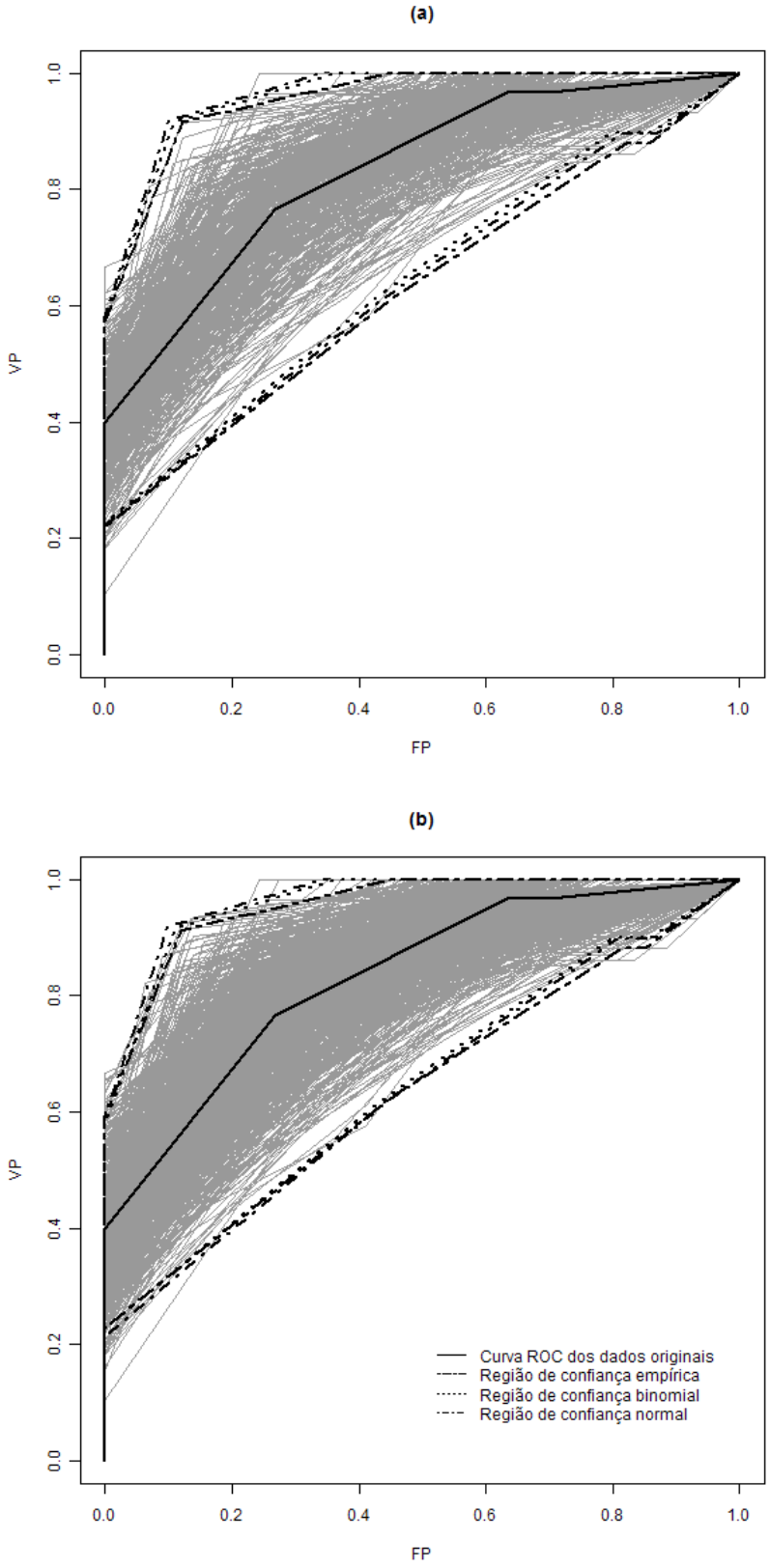

FIGURA 6.18: Bandas de confiança estimadas pelo método MLO para: (a) 500 reamostragens e (b) 1000 reamostragens.

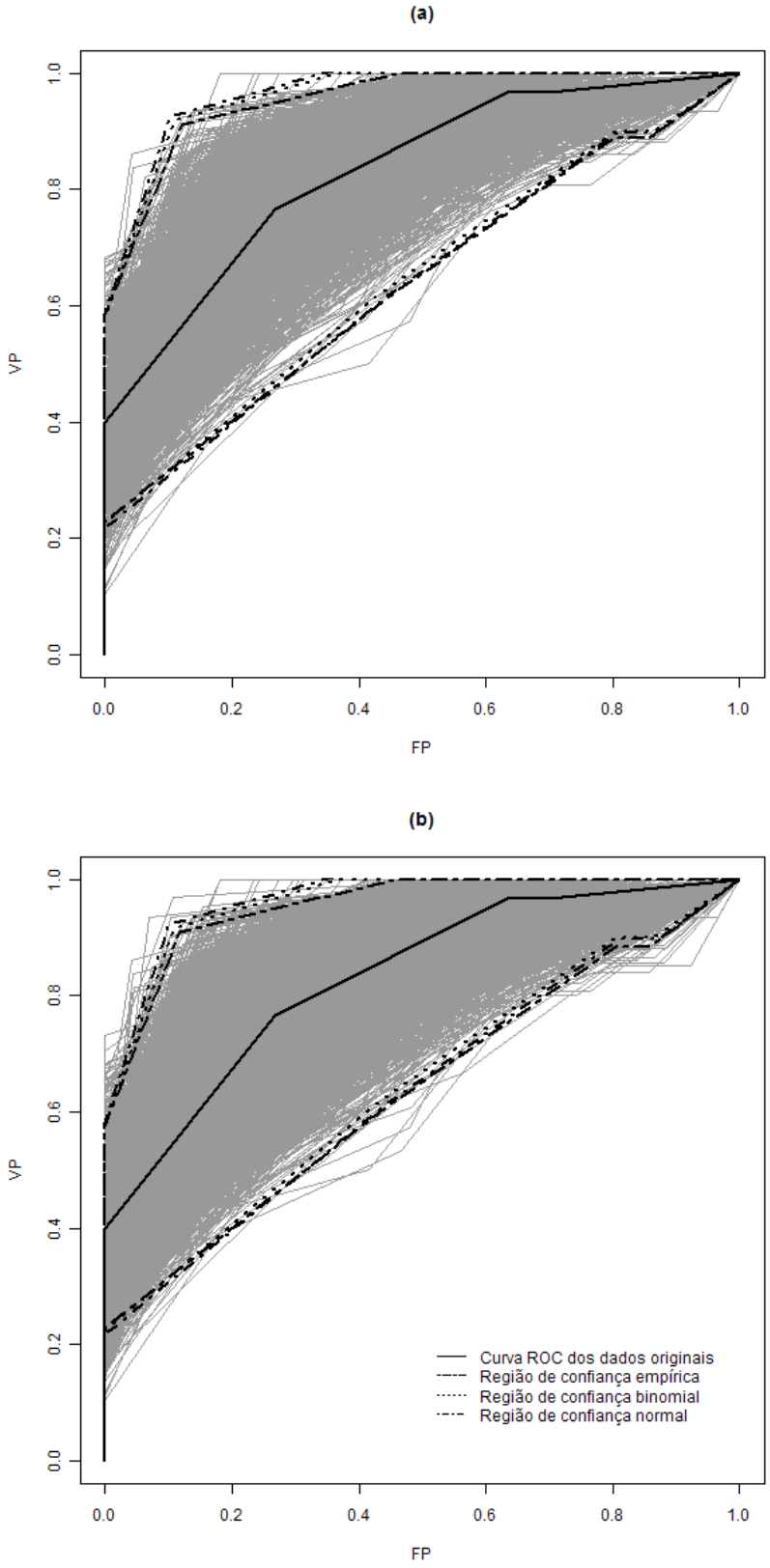

FIGURA 6.19: Bandas de confiança estimadas pelo método MLO para: (a) 2500 reamostragens e (b) 5000 reamostragens.

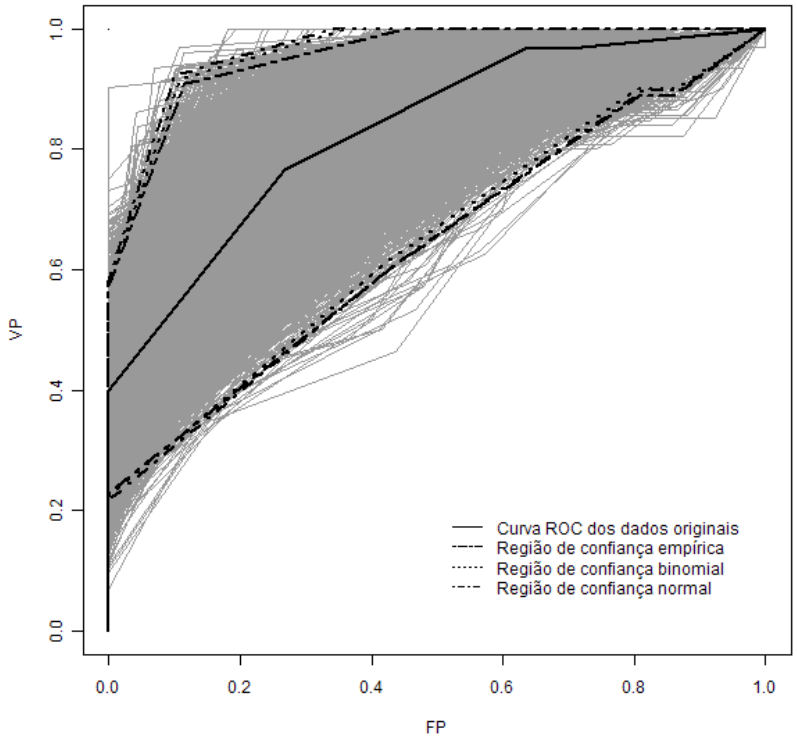

FIGURA 6.20: Bandas de confiança estimadas pelo método MLO para 10000 reamostragens.

Após realizadas todas as B reamostrangens *bootstrap* e obtidas as curvas ROC para cada reamostragem, encontramos a proporção das curvas ROC que estão completamente dentro das bandas de confiança estimadas pelo método MLO, para cada uma das distribuições assumidas. Na Tabela 6.7 dispomos os valores da probabilidade de cobertura simulada por esse método.

TABELA 6.7: Probabilidade de cobertura das bandas de confiança MLO

	Número de reamostragens *bootstrap*						
Probabilidade de cobertura	**50**	**100**	**500**	**1000**	**2500**	**5000**	**10000**
Empírica	0.98	0.97	0.928	0.929	0.9248	0.9268	0.9261
Binomial	0.96	0.96	0.926	0.924	0.932	0.9346	0.9357
Normal	0.96	0.97	0.934	0.941	0.9484	0.9468	0.9463

Na Figura 6.21, com os valores dispostos na Tabela 6.7, construímos o gráfico das probabilidades de cobertura simulada *versus* a quantidade de reamostragens *bootstrap*.

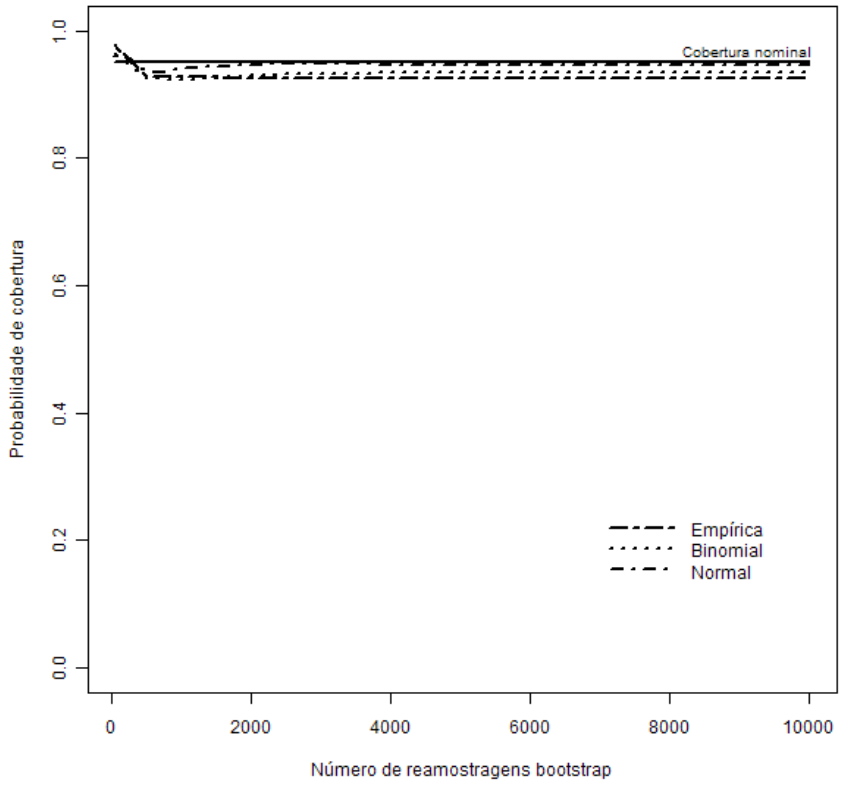

FIGURA 6.21: Probabilidades de cobertura para o método MLO.

Analisando a Tabela 6.7 e a Figura 6.21 notamos que o valor da probabilidade de cobertura simulada, apesar de ser muito próximo ao valor da probabilidade de cobertura nominal, é inferior a 0.95, para as três distribuições assumidas na geração de intervalos pelo método MLO. Entretanto, de todos os métodos de estimação de bandas pontuais até então apresentados, este é o método que mais cobre as curvas *ROC* geradas pelo *bootstrap*.

Para bandas geradas pela distribuição empírica, quando efetuado um número reduzido de reamostragens, a cobertura é em torno de 97% das curvas *ROC* geradas. Entretanto, conforme o número de reamostragens aumenta, a cobertura de tais bandas diminui e fica em torno de 92.5% do total das curvas geradas. Também notamos que para a distribuição empírica, a probabilidade de cobertura simulada é a menor de todas as coberturas estimadas, quando

comparado com as outras duas distribuições assumidas.

Nas bandas geradas pela distribuição binomial, quando efetuado um número reduzido de reamostragens, a cobertura é em torno de 96% das curvas *ROC* geradas. Entretanto, conforme o número de reamostragens aumenta, a cobertura de tais bandas diminui e fica em torno de 93.5% do total das curvas geradas.

Para bandas geradas pela distribuição normal, quando efetuado um número reduzido de reamostragens, a cobertura é em torno de 97% das curvas *ROC* geradas. Entretanto, conforme o número de reamostragens aumenta, a cobertura de tais bandas diminui e fica em torno de 94.6% do total das curvas geradas. Este é o método que mais cobre as curvas *ROC* reamostradas e também é o mais estável, pois o valor da cobertura das bandas normais pouco oscila conforme varia o número de reamostragens *bootstrap*.

Bandas de confiança para o método das juntas simultâneas (SJR)

Macskassy e Provost (2004) propõem usar intervalos normais para encontrar as bandas de confiança para a curva *ROC* estimada pelo método SJR. Entretanto, como este é um método de geração não paramétrico (*bootstrap* não paramétrico), achamos prudente encontrar tais bandas através de intervalos percentis, usando assim a distribuição empírica.

Os intervalos *bootstrap* percentis foram gerados de tal forma que os $\frac{\alpha}{2}$-ésimo e $(1-\frac{\alpha}{2})$-ésimo quantis da distribuição empírica das curvas *ROC* reamostradas, organizados em forma crescente, fossem estimados.

Baseado no valor da estatística *KS* apresentado na Tabela B.1, encontramos para o nível de significância $\alpha = 0.05$, as seguintes distâncias

$$g = \frac{1.36}{\sqrt{n}} = \frac{1.36}{\sqrt{30}} = 0.2483009,$$

e

$$h = \frac{1.36}{\sqrt{m}} = \frac{1.36}{\sqrt{30}} = 0.2483009,$$

em que g é a distância horizontal e h é a distância vertical.

Nas Figuras 6.22, 6.23, 6.24 e 6.25 dispomos as curvas *ROC* geradas por reamostragem *bootstrap* para o método SJR, bem como suas bandas de confiança estimadas pela distribuição empírica.

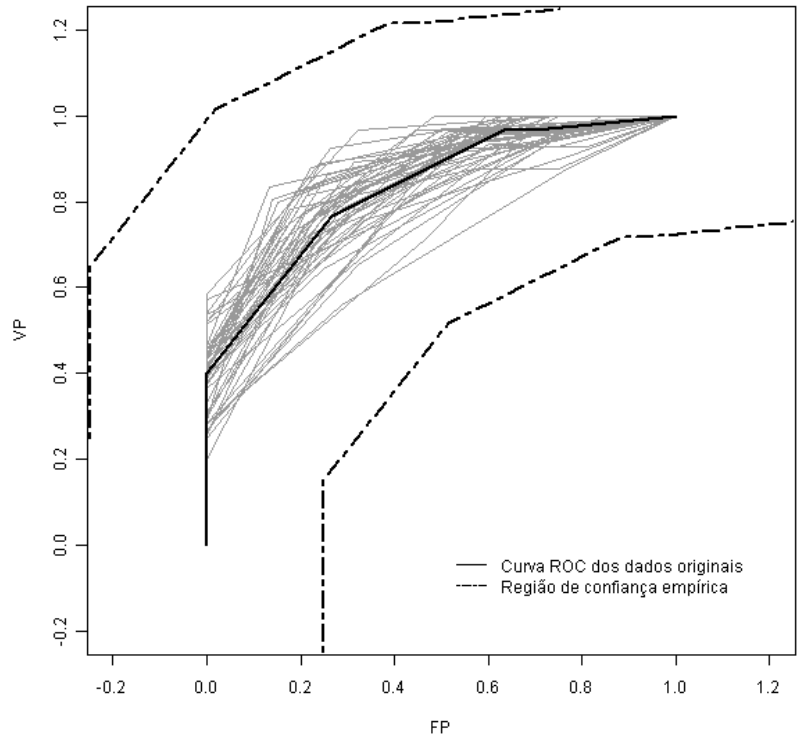

FIGURA 6.22: Bandas de confiança estimadas pelo método SJR para 50 reamostragens.

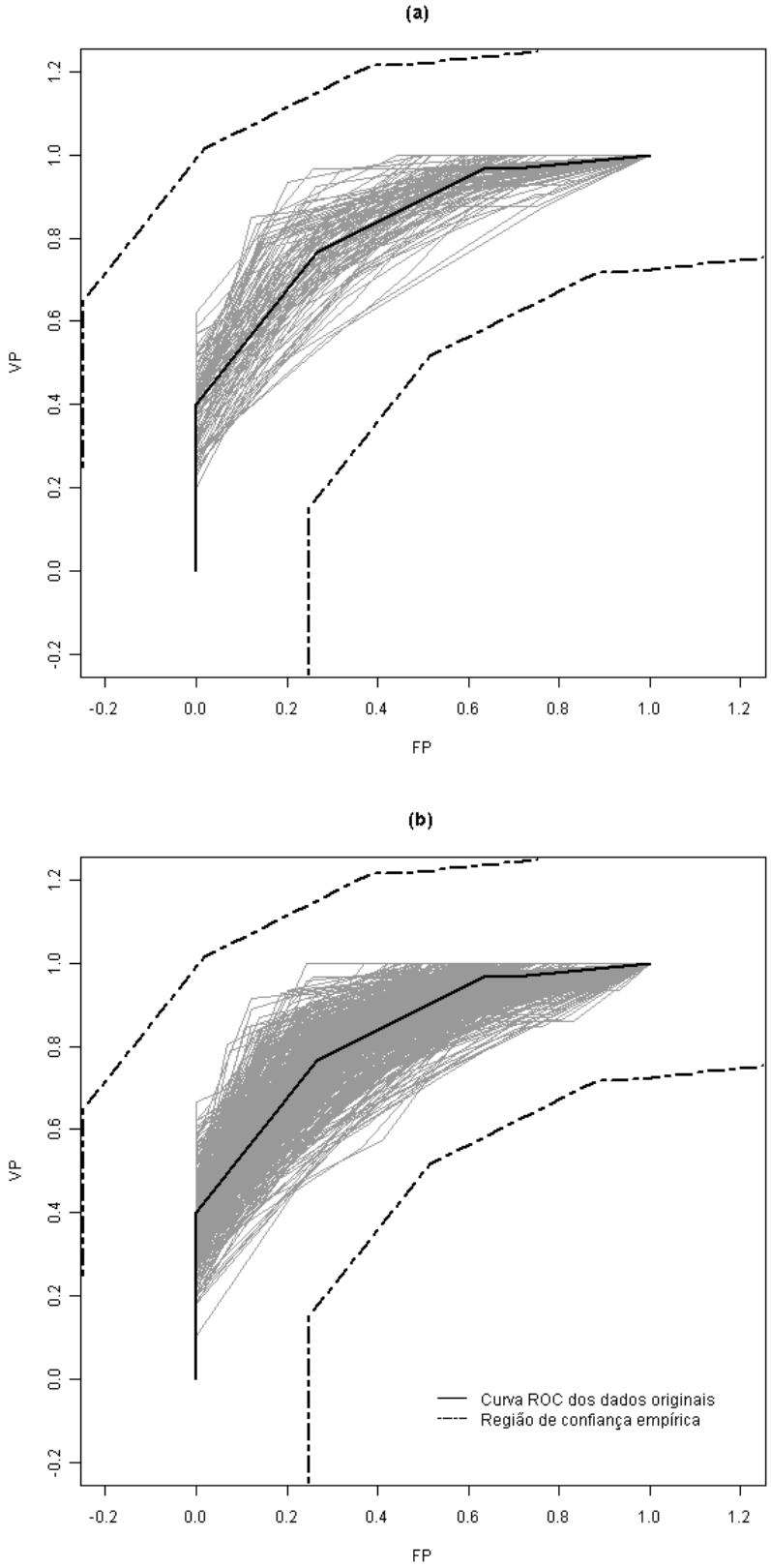

FIGURA 6.23: Bandas de confiança estimadas pelo método SJR para: (a) 100 reamostragens e (b) 500 reamostragens.

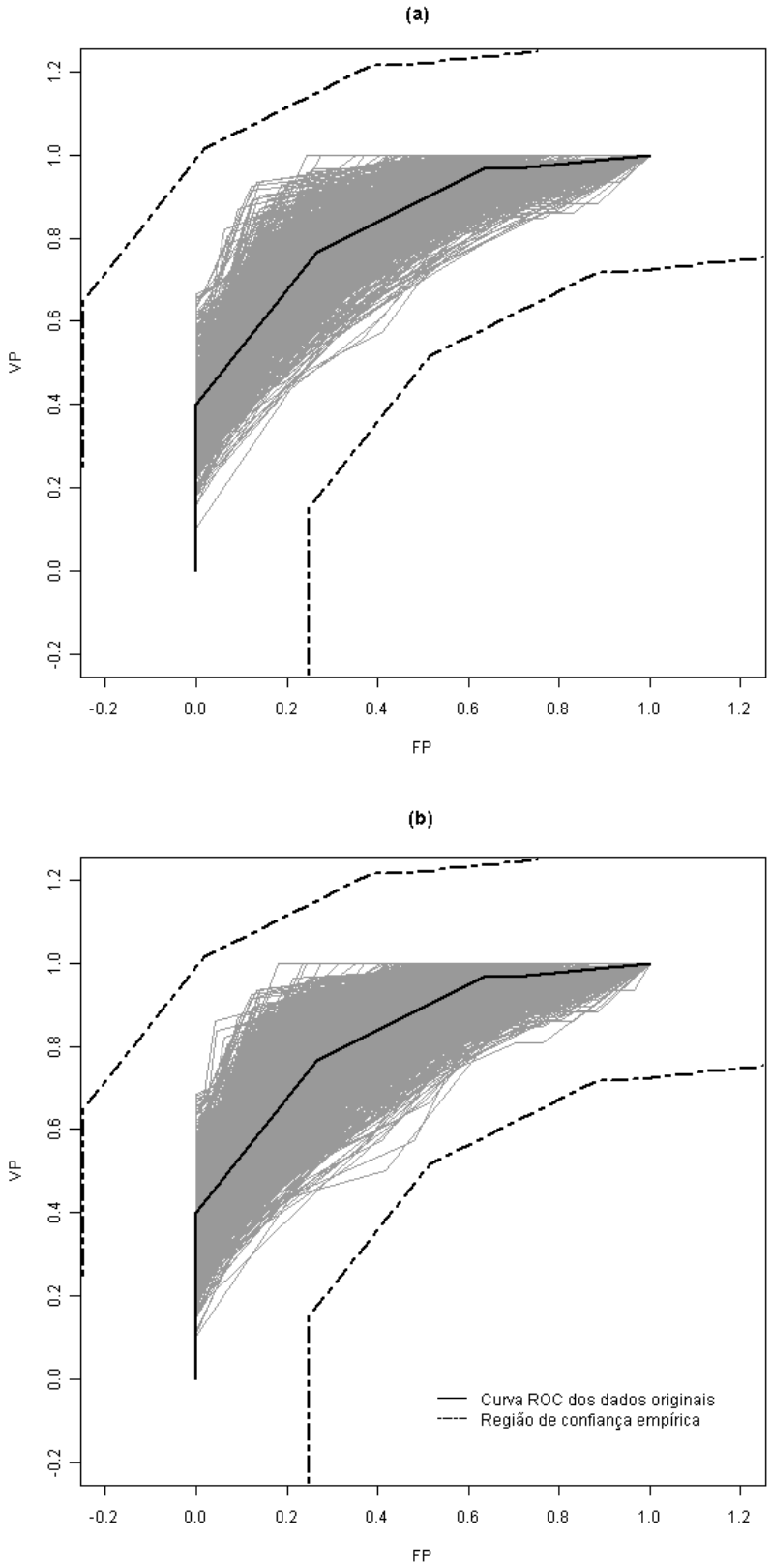

FIGURA 6.24: Bandas de confiança estimadas pelo método SJR para: (a) 1000 reamostragens e (b) 2500 reamostragens.

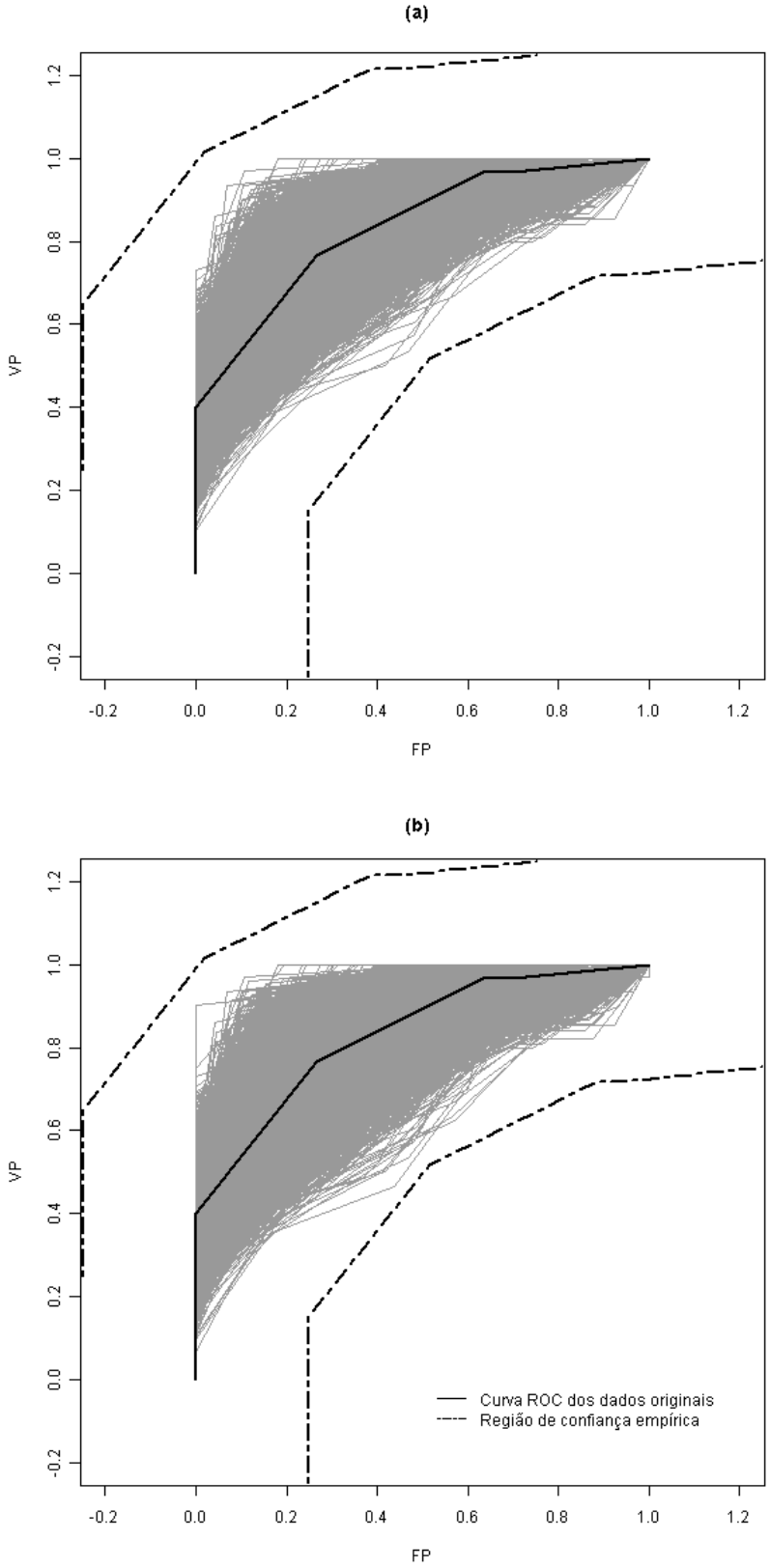

FIGURA 6.25: Bandas de confiança estimadas pelo método SJR para: (a) 5000 reamostragens e (b) 10000 reamostragens.

Após realizadas todas as B reamostrangens *bootstrap* e obtidas as curvas ROC para cada reamostragem, encontramos a proporção das curvas ROC que estão completamente dentro das bandas de confiança estimadas pelo método SJR, para a distribuição assumida. Na Tabela 6.8 dispomos os valores da probabilidade de cobertura simulada por esse método.

TABELA 6.8: Probabilidade de cobertura das bandas de confiança SJR

	Número de reamostragens *bootstrap*						
	50	**100**	**500**	**1000**	**2500**	**5000**	**10000**
Probab. cobertura empírica	1	1	1	1	1	1	1

Como o nosso conjunto de dados é constituído da mesma quantidade de indivíduos doentes e não doentes ($m = 30 = n$), os valores das distâncias horizontais e verticais são iguais. Assim, não encontramos retângulos, mas quadrados. Logo, cada ponto de corte empírico pertencente à curva ROC estimada é o ponto central de um quadrado de lado $2g = 0.4966018 = 2h$.

Conforme a descrição deste método no Capítulo 5, a distância entre as fronteiras superior e inferior é a diagonal do quadrado de lado 0.4966018, que é dada por

$$\zeta^2 = 0.4966018^2 + 0.4966018^2$$

$$\zeta = \sqrt{0.4966018^2 + 0.4966018^2}$$

$$\zeta = 0.702301.$$

Quando comparamos o valor 0.702301 com o valor da diagonal de um quadrado unitário (1.414214) em que qualquer curva ROC esta delimitada, observamos que o valor 0.702301 é aproximadamente metade do valor 1.414214. Logo, as bandas de confiança empíricas para o método SJR são relativamente largas para este específico conjunto de dados.

Analisando as Figuras 6.22, 6.23, 6.24 e 6.25, a Tabela 6.8 e as observações anteriores, independentemente do número de reamostragens *bootstrap*, este método de geração de bandas de confiança regionais tem uma de cobertura simulada de 100%, valor superior ao da cobertura esperada que é de $(95\%)^2$.

Se as amostras de indivíduos doentes e não doentes fossem de quantidades maiores e de diferentes valores, teríamos menores distâncias g e h, tendo assim retângulos com menores dimensões, podendo não conter todas as curvas ROC geradas por reamostragem, ou seja, a probabilidade de cobertura de tais bandas teria um valor menor do que o encontrado em nosso exemplo.

Bandas de confiança para o método das larguras fixas (FWB)

A forma de gerar intervalos de confiança pelo método FWB é diferente das anteriores, pois desejamos encontrar uma região de confiança global para a curva ROC estimada.

Os intervalos *bootstrap* percentis foram gerados de tal forma que os $\frac{\alpha}{2}$-ésimo e $(1-\frac{\alpha}{2})$-ésimo quantis da distribuição empírica das curvas ROC reamostradas, organizados em forma crescente, fossem estimados.

Este método de geração global da região de incerteza associada à curva ROC estimada busca encontrar uma distância $2d$ que cubra as curvas ROC reamostradas pelo *bootstrap* com uma probabilidade de cobertura simulada de igual valor a probabilidade de cobertura nominal, ou seja, as bandas de confiança globais devem conter 95% das curvas reamostradas.

Considerando a inclinação sugerida por Campbell (1994), que é de

$$\varrho = -\sqrt{\frac{n}{m}} = -\sqrt{\frac{30}{30}} = -1,$$

as fronteiras superior e inferior para a região de confiança em torno da curva ROC são construídas aumentando, sucessivamente, 0.001 nos valores de $2d$, até encontrar um valor final para $2d$ cuja probabilidade de cobertura seja de 95% do total de curvas ROC reamostradas.

Nas Figuras 6.26, 6.27, 6.28 e 6.29 dispomos as curvas ROC geradas por reamostragem *bootstrap* para o método FWB, bem como suas bandas de confiança estimadas pela distribuição empírica.

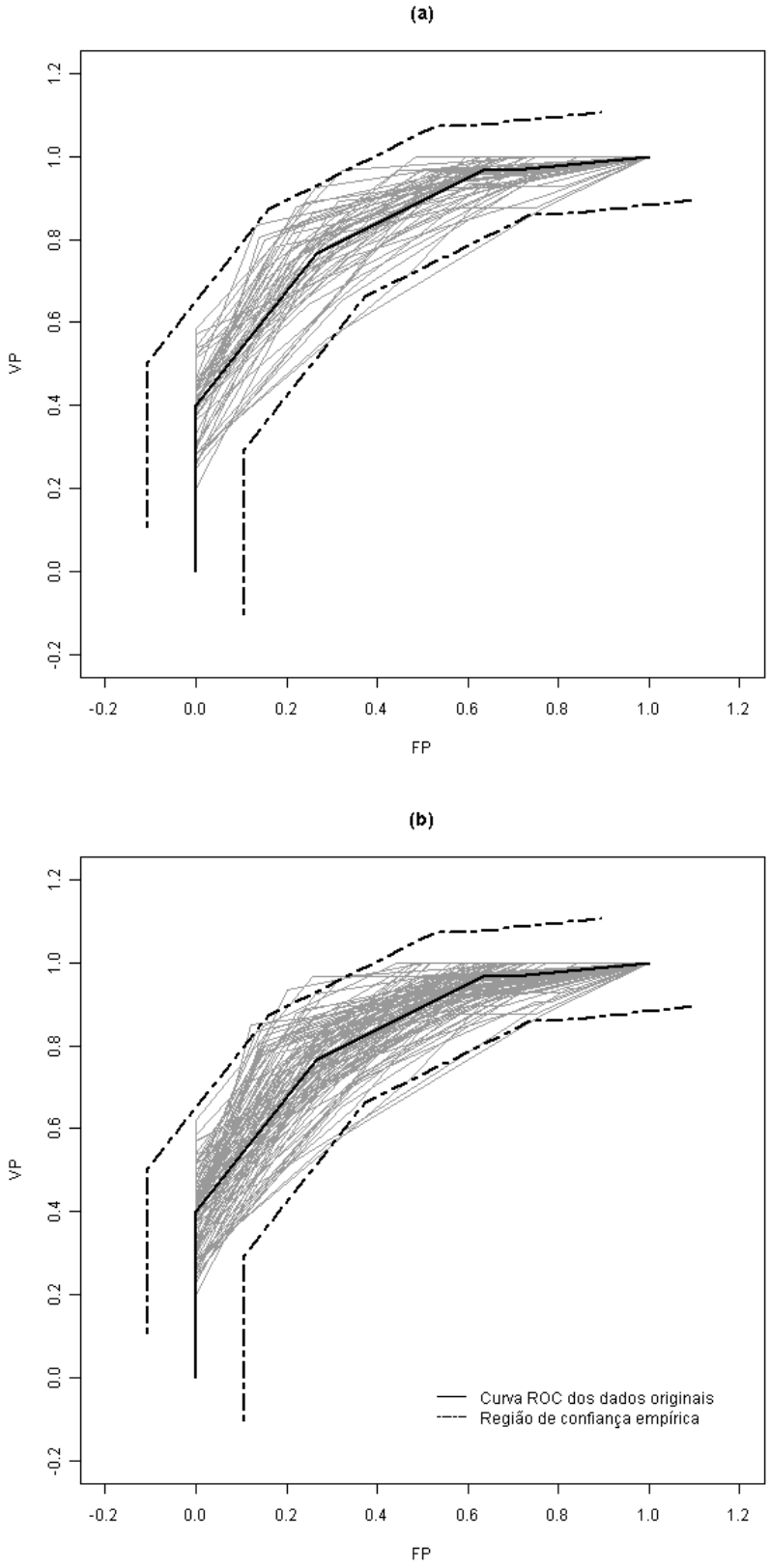

FIGURA 6.26: Bandas de confiança estimadas pelo método FWB para: (a) 50 reamostragens e (b) 100 reamostragens.

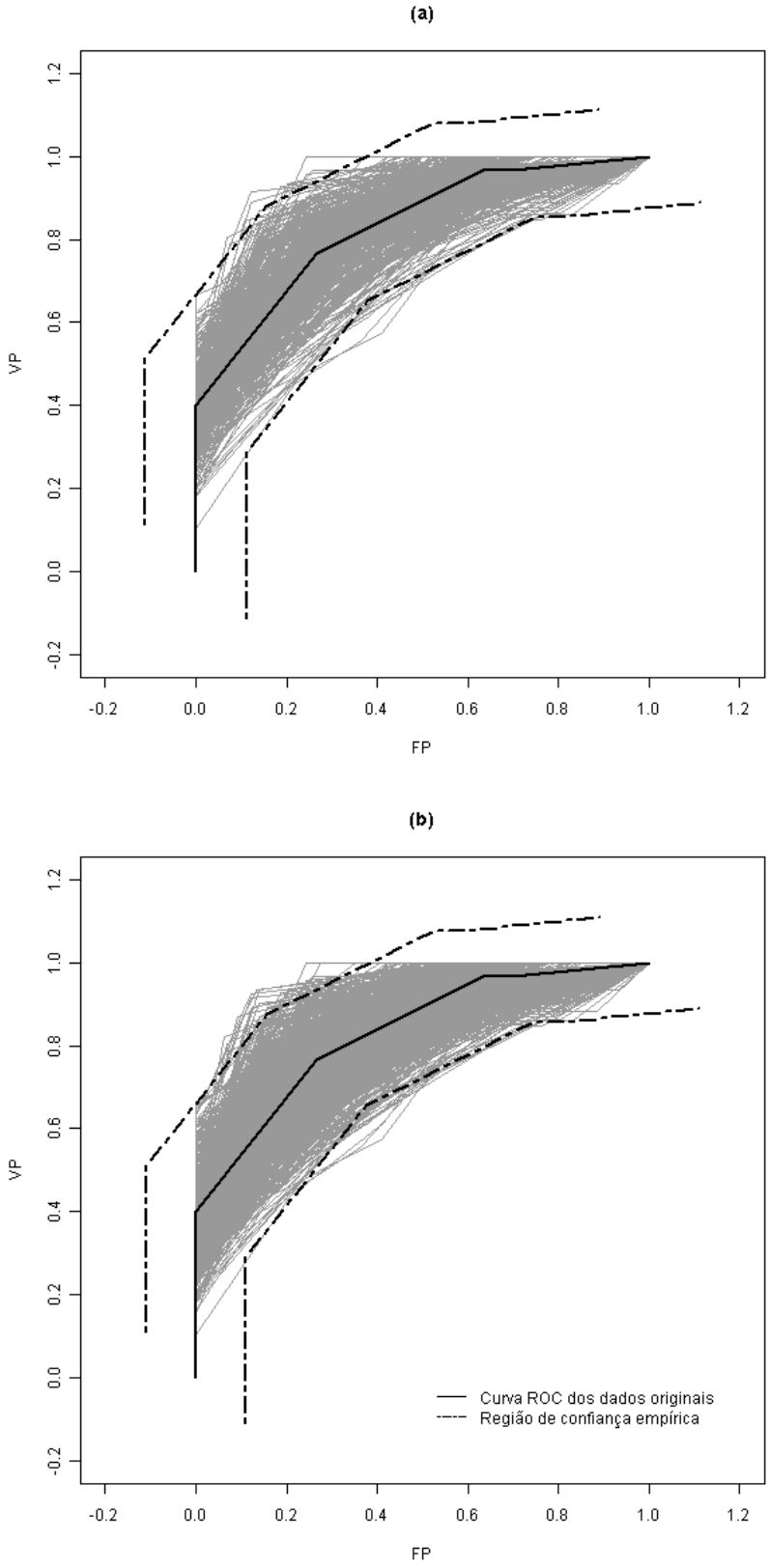

FIGURA 6.27: Bandas de confiança estimadas pelo método FWB para: (a) 500 reamostragens e (b) 1000 reamostragens.

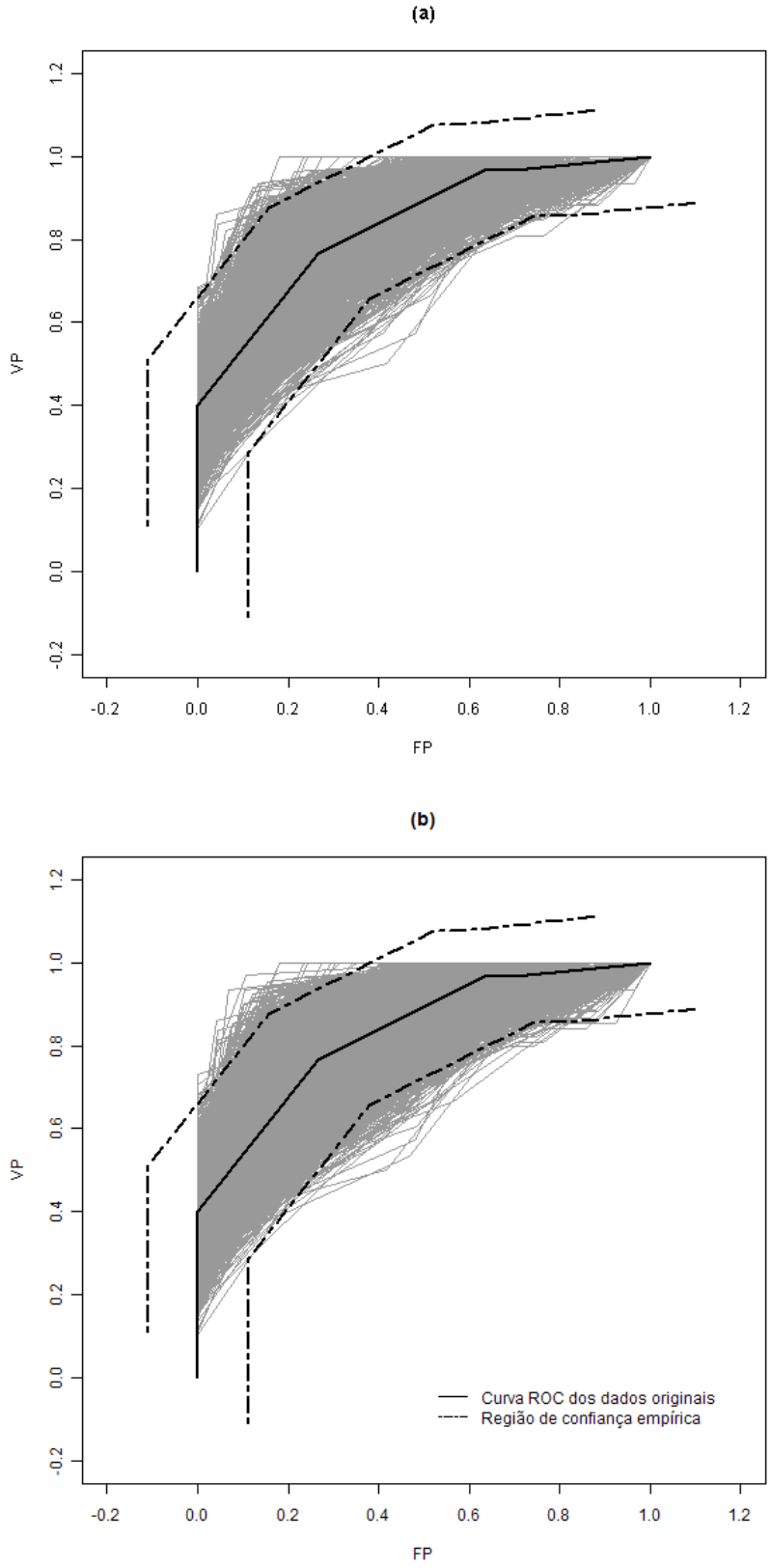

FIGURA 6.28: Bandas de confiança estimadas pelo método FWB para: (a) 2500 reamostragens e (b) 5000 reamostragens.

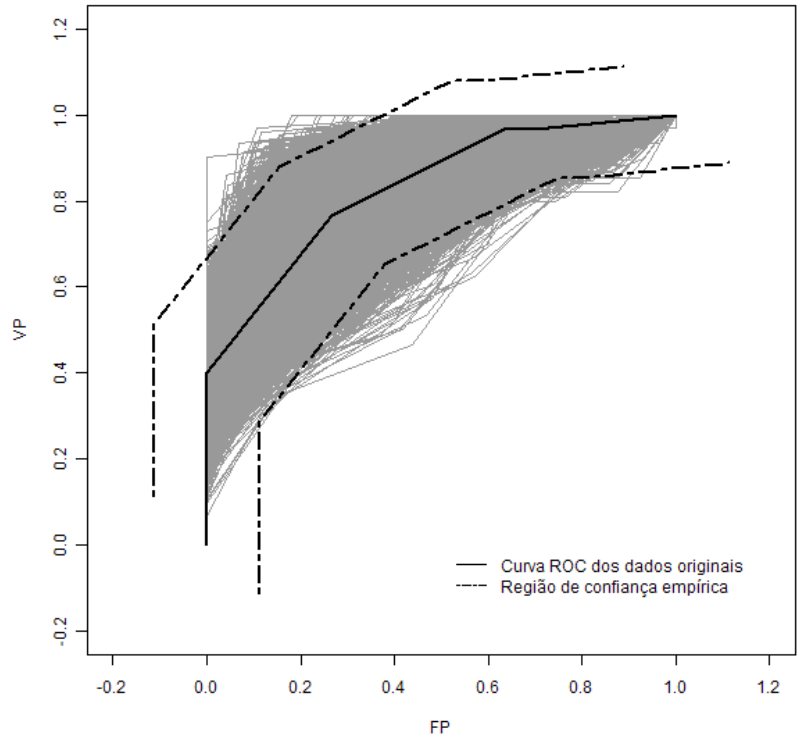

FIGURA 6.29: Bandas de confiança estimadas pelo método FWB para 10000 reamostragens.

Nas Tabelas 6.9 e 6.10 dispomos os quatro últimos valores de $2d$ e a probabilidade de cobertura simulada para os últimos quatro incrementos executados.

TABELA 6.9: Probabilidade de cobertura das bandas de confiança FWB para os últimos 4 incrementos em 50, 100, 500 e 1000 reamostragens

	Número de reamostragens *bootstrap*			
	50	**100**	**500**	**1000**
distância $2d$	0.146	0.146	0.156	0.153
Prob. de cobertura para $2d$	0.94	0.93	0.946	0.942
distância $2d$	0.147	0.147	0.157	0.154
Prob. de cobertura para $2d$	0.94	0.93	0.946	0.945
distância $2d$	0.148	0.148	0.158	0.155
Prob. de cobertura para $2d$	0.94	0.93	0.946	0.948
distância $2d$	**0.149**	**0.149**	**0.159**	**0.156**
Prob. de cobertura para $2d$	0.98	0.95	0.95	0.95

TABELA 6.10: Probabilidade de cobertura das bandas de confiança FWB para os últimos 4 incrementos em 2500, 5000 e 10000 reamostragens

	Número de reamostragens *bootstrap*		
	2500	**5000**	**10000**
distância $2d$	0.154	0.154	0.156
Prob. de cobertura para $2d$	0.9432	0.9446	0.94861
distância $2d$	0.155	0.155	0.157
Prob. de cobertura para $2d$	0.9000	0.9470	0.9472
distância $2d$	0.156	0.156	0.158
Prob. de cobertura para $2d$	0.9492	0.9494	0.9487
distância $2d$	**0.157**	**0.157**	**0.159**
Prob. de cobertura para $2d$	0.9504	0.9504	0.9515

Analisando a Tabela 6.9 para um pequeno valor de reamostragens *bootstrap*, por exemplo 50, a probabilidade de cobertura relacionada ao penúltimo incremento ($2d = 0.148$) é de 94% das curvas ROC reamostradas, contra a cobertura de 98% do total de curvas geradas para a distância final ($2d = 0.149$).

Analisando a Tabela 6.10 para um grande valor de reamostragens *bootstrap*, por exemplo 10000, a probabilidade de cobertura relacionada ao penúltimo incremento ($2d = 0.158$) é de 94.87% das curvas ROC reamostradas, contra a cobertura de 95.15% do total de curvas geradas para a distância final ($2d = 0.159$).

Logo, concluímos que o valor da distância $2d$ encontra-se, independentemente do número de reamostragens, no intervalo [0.149 ; 0.159], porém com uma maior frequência em torno do limite superior do mesmo. Concluímos também que para uma pequena variação da distância $2d$, quando o número de reamostragens é reduzido, há uma maior variação na probabilidade de cobertura das bandas, enquanto que para um grande número de reamostragens, há pouca variação na probabilidade de cobertura das bandas FWB.

Capítulo 7

Considerações Finais

Neste trabalho, revisamos, elaboramos e reproduzimos algumas técnicas conhecidas para encontrar bandas de confiança para a curva *ROC* estimada a partir de um teste diagnóstico.

Para um melhor entendimento do assunto, no Capítulo 2 revisamos sobre testes diagnósticos; no Capítulo 3 estudamos a curva *ROC* para testes diagnósticos; no Capítulo 4 revisamos o método de reamostragem *bootstrap*; no Capítulo 5 revisamos e propomos algumas técnicas de geração de regiões de incerteza para a curva *ROC*, para finalmente, aplicarmos e reproduzirmos essas técnicas no Capítulo 6.

No Capítulo 6, para a região de incerteza encontrada por bandas de confiança pontuais através dos métodos VA, MO, TA e MLO, independentemente do número de reamostragens *bootstrap*, a probabilidade de cobertura simulada subestima a probabilidade de cobertura nominal. Para a região de incerteza encontrada por bandas de confiança regionais através do método SJR, obtemos independentemente do número de reamostragens *bootstrap*, uma probabilidade de cobertura simulada de 100%, valor este que sobrestima a cobertura esperada. Para a região de incerteza encontrada por bandas de confiança globais através do método FWB, utilizamos para percorrer as taxas de FP, uma inclinação de valor -1, e encontramos uma distância $2d$, compreendida no intervalo $[0.149\ ;\ 0.159]$, o qual resulta numa cobertura de 95% de todas as curvas *ROC* geradas por

reamostragem *bootstrap*.

Quando comparado os dois métodos para geração de bandas de confiança pontuais verticais, portanto unidimensionais, podemos notar que o método VA tem um melhor desempenho em relação ao método TA, pois o primeiro cobre uma proporção maior de curvas *ROC* geradas por reamostragem *bootstrap*. Contudo, quando se trata de intervalos unidimensionais combinados, ou seja, intervalos encontrados tanto no eixo das abscissas quanto no eixo das ordenadas, o método MLO tem um melhor desempenho se comparado com o método MO, pois as bandas de confiança encontradas a partir da geração do método MLO apresentam uma probabilidade de cobertura maior do que a das bandas encontradas pelo método MO.

Em geral, das quatro maneiras apresentadas neste trabalho para encontrar regiões de confiança pontuais para a curva *ROC*, para este específico conjunto de dados, o método MLO tem um melhor desempenho, pois sua probabilidade de cobertura simulada é mais próxima do valor nominal. Já o método MO, quando comparado todos os quatro métodos de geração de bandas pontuais e quando utilizadas as três distribuições para geração dos intervalos de confiança, tem um pior desempenho, pois a cobertura das bandas encontradas pela distribuição empírica e pela distribuição binomial é baixa e tem valor bem distante da cobertura nominal.

Notamos que a probabilidade de cobertura para os quatro métodos de geração de bandas pontuais se estabiliza para um número superior a 1000 reamostragens *bootstrap*. Percebemos também que quanto maior for o número de reamostragens, mais a probabilidade de cobertura vai diminuindo proporcionalmente, ou seja, há um decaimento dos valores da probabilidade de cobertura. Uma possível explicação para isso é devido à rigidez do método adotado, já que se apenas um ponto de qualquer curva reamostrada cair fora das fronteiras estabelecidas, desconsideramos por completo essa curva no instante da contagem da cobertura.

Como proposta futura, pretendemos elaborar outras formas de encontrar regiões de incerteza pontuais, globais e regionais para a curva *ROC*.

Referências

[1] ALTMAN , D. G. *Practical statistics for medical research*. 1.ed. London: Chapman & Hall, 1991.

[2] ALTMAN, D. G.; BLAND, J. M. Diagnostic tests 3: Receiver operating characteristic plots. **British Medical Journal**, v.309, n.6948, p.188, July, 1994.

[3] BAMBER, D. The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. **Journal of Mathematical Psychology**, v.12, n.4, p.387-415, November, 1975.

[4] BROEMELING, L. D. *Bayesian Biostatistics and Diagnostic Medicine*. 1.ed. Chapman & Hall / CRC, 2007.

[5] CAMPBELL, G. Advances in statistical methodology for the evaluation of diagnostic and laboratory tests. **Statistics in Medicine**, v.13, n.5/7, p.499-508, March/April, 1994.

[6] CANTY, A. *The bootstrap and confidence intervals*. Disponível em: <http://mathstat.concordia.ca/canty/teaching.mast679t.html> Acesso em 03 de Março de 2009.

[7] COLLINSON, P. Of bombers, radiologists, and cardiologists: Time to ROC. **Heart**, v.80, n.3, p.215-217, September, 1998.

[8] CONOVER, W. J. *Practical Nonparametric Statistics.* New York: Wiley, 1.ed, 1971.

[9] DUNN, G.; EVERITT, B. *Clinical biostatistics: An introduction to evidence-based medicine*. London: Edward Arnold, 1995.

[10] EFRON, B. Bootstrap methods: Another look at the jackknife. **The Annals of Statistics**, v.7, n.1, p.1-26, January, 1979.

[11] EFRON, B.; TIBSHIRANI, R. *An introduction to the bootstrap*. New York: Chapman & Hall, 1993.

[12] EGAN, J. P. *Sigmal detection theory and ROC analysis*. New York: Academic Press, 1975.

[13] EVANS, A. L . Medical Physics Handbooks 10: *The Evaluation of Medical Images*. Adam Hilger Ltd., Bristol, England, 1981.

[14] FAWCETT, T. ROC Graphs: Notes and Practical Considerations for Data Mining Researchers. Technical Report HPL-2003-4, HP Labs, January, 2003.

[15] FLEISS, J. L. *Statistical methods for rates and proportions*. 2.ed. New York: John Wiley, 1981.

[16] FLETCHER, R.; FLETCHER, S.; WAGNER, E. *Epidemiologia clínica: Elementos essenciais*. Trad. Bruce B. Duncan, Maria I. Schmidt. 3.ed. Porto Alegre: Artes Médicas, 1996.

[17] HILGERS, R. A. Distribution-free confidence bounds for ROC curves. **Methods of Information in Medicine**, v.30, n.2, p.96-101, April, 1991.

[18] JENSEN, K.; MÜLLER, H. H.; SCHÄFER, H. Regional confidence bands for ROC curves. **Statistics in Medicine**, v.19, n.4, p.493-509, February, 2000.

[19] LINNET, K. Comparison of quantitative diagnostic tests: Type I error, power, and sample size. **Statistics in Medicine**, v.6, n.2, p.147-158, March, 1987.

[20] LINNET, K. A review on the methodology for assessing diagnostic tests. **Clinical Chemistry**, v.34, n.7, p.1379-1386, July, 1988.

[21] MACSKASSY, S. A.; PROVOST, F. Confidence Bands for ROC Curves: Methods and an Empirical Study. **Appears in First Workshop on ROC Analysis in AI**, ECAI-2004, Spain, 2004.

[22] MACSKASSY, S. A.; PROVOST, F.; LITTMAN, M. L. Confidence Bands for ROC Curves. Disponível em: <http://www.research.rutgers.edu/∼sofmac/paper/ceder2003/macskassy-CeDER-is-03-04.pdf> Acesso em 07 de Janeiro de 2008.

[23] MARTINEZ, E. Z. *Métodos estatísticos para estudos de desempenho de Testes Diagnósticos*. Dissertação de Mestrado em Estatística - Universidade Federal de São Carlos, 2001.

[24] MARTINEZ, E. Z.; LOUZADA-NETO, F.; PEREIRA, B. B. A curva ROC para Testes Diagnósticos. **Cadernos Saúde Coletiva**, Rio de Janeiro, v.11, p.7-31, 2003.

[25] MARUSSI, E. F. *Análise de morfologia ultra-sonográfica aliada à colordopplervelocimetria na previsão do diagnóstico histológico dos nódulos sólidos da mama*. Tese de Doutorado em Medicina - Faculdade de Ciências Médicas, Universidade Estadual de Campinas, 2001.

[26] METZ, C. E. Statistical analysis of ROC data in evaluating diagnostic performance. **Multiple Regression Analysis: Applications in the Health Sciencs**, n.13, p.365-384, 1986.

[27] REISER, B.; FARAGGI, D. Confidence intervals for generalized ROC criterion. **Biometrics**, v.53, p.644-652, June, 1997.

[28] RIEGELMAN, R. K.; HIRSCH, R. P. *Studying a study and testing a test: How to read the health science literature.* 3.ed. Boston: Little, Brow and Company, 1996.

[29] SARAIVA, K. F. *Inferência Bayesiana para Teste Diagnóstico*. Dissertação de Mestrado em Estatística - Universidade Federal de São Carlos, 2004.

[30] SCHÄFER, H. Efficient confidence bounds for ROC curves. **Statistics in Medicine**, v.13, n.15, p.1551-1561, August, 1994.

[31] SOX, H. C. Probability theory in the use of diagnostic test: An introduction to critical study of the literature. **Annals of Internal Medicine**, v.104, n.1, p.60-66, January, 1986.

[32] SWETS, J. A. *Signal Detection Theory and ROC Analysis in Psychology and Diagnostics*. Collected Papers. New Jersey: LEA, 1996.

[33] TACONELI, C. A. *Reamostragem bootstrap em amostragem por conjuntos ordenados e intervalos de confiança não paramétricos para a média*. Dissertação de Mestrado em Estatística - Universidade Federal de São Carlos, 2005.

[34] TACONELI, C. A.; BARRETO, M. C. Avaliação de uma proposta de intervalos de confiança bootstrap em amostragem por conjuntos ordenados perfeitamente. **Revista de Matemática e Estatística**, São Paulo, v.23, n.3, p.33-53, 2005.

[35] THIBODEAU, L. A. Evaluating diagnostic tests. **Biometrics**, v.37, p.801-804, December, 1981.

[36] ZHOU, X. H.; OBUCHOWSKI, N. A.; MCCLISH, D. K. *Statistical Methods in Diagnostic Medicine*. Copyright by John Wiley & Sons, Inc., New York, 2002.

Apêndice A

Curva *ROC* no Plano Binormal

Uma curva ROC é a representação gráfica dos pares S_E (ou TVP) no eixo das ordenadas e $1 - E_S$ (ou TFP) no eixo das abscissas, resultantes da variação do valor de corte ao longo de um eixo de decisão. A representação gráfica assim resultante é conhecida como curva ROC no plano unitário.

Segundo Metz (1986), uma curva ROC é uma descrição empírica da capacidade do sistema de diagnóstico poder discriminar entre dois estados, num universo onde cada ponto da curva representa um compromisso diferente entre a S_E e a $1 - E_S$, ponto este adquirido por adotar diferentes valores de corte de anormalidade ou nível crítico de confiança no processo de decisão.

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

A forma de representar os dados da curva ROC no plano binormal é através do papel de probabilidade normal. Seja S uma variável aleatória contínua que representa o resultado de um teste diagnóstico. Segundo Metz (1986), a escala de probabilidade é construída calculando o valor z, para cada valor de S, de acordo com a seguinte equação

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

Logo, o papel de probabilidade normal usa escalas de probabilidade para cada um dos seus eixos (ver Figura A.1).

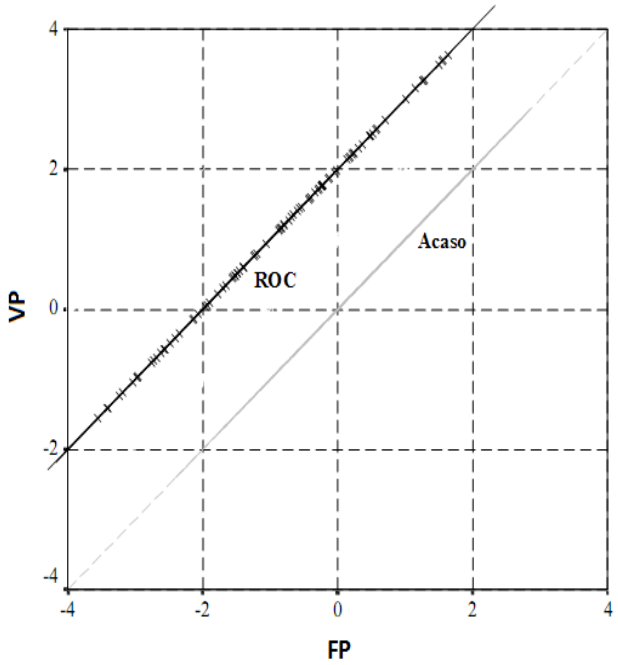

FIGURA A.1: Representação da curva *ROC* no plano binormal.

Uma vantagem desse tipo de transformação é que a curva *ROC* para distribuições normais é uma reta e a separação entre a médias das duas distribuições pode ser retirada do próprio gráfico como função da diferença entre o valor da ordenada e o valor da abscissa.

Metz (1986) relata que "de um modo geral, uma curva *ROC* é especificada assumindo uma forma particular com um ou mais parâmetros ajustáveis (...). A forma funcional binormal para a curva *ROC* é utilizada muito frequentemente e verifica que esta forma fornece bons ajustes às curvas *ROC* empíricas, medidas numa grande variedade de situações".

Swets (1996) menciona que "nem todos os dados no plano binormal se ajustam a uma reta (...) e que um desvio à linearidade viola a hipótese de normalidade, e um declive não unitário viola a hipótese de igualdade das variâncias".

Segundo Metz (1986), a curva *ROC* binormal pode ser interpretada, em

termos da variável de decisão S, proveniente de duas densidades Gaussianas, por

$$f(s \mid h_1) = \frac{1}{\sqrt{2\pi\sigma_D^2}} \ exp\left(-\frac{(s-\mu_D)^2}{2\sigma_D^2}\right),$$

em que $f(s \mid h_1)$ corresponde à função densidade de probabilidade para os casos dos indivíduos doentes, e

$$f(s \mid h_0) = \frac{1}{\sqrt{2\pi\sigma_{\overline{D}}^2}} \ exp\left(-\frac{(s-\mu_{\overline{D}})^2}{2\sigma_{\overline{D}}^2}\right),$$

em que $f(s \mid h_0)$ corresponde à função densidade de probabilidade para os casos dos indivíduos não doentes.

Por comodidade, denotaremos por D o paciente doente (D = 1) e por $\overline{D}$ o paciente não doente (D = 0).

Sejam $S_{\overline{D}} \sim N(0\ ,\ 1)$ e $S_D \sim N\left(\frac{a}{c}\ ,\ \frac{1}{c}\right)$.

Logo, a forma funcional binormal para a curva ROC pode ser expressa pelo par de equações

$$(S_E)(r) = \Phi(a - cr)$$

e

$$(1 - E_S)(r) = \Phi(-r), \tag{A.1}$$

no qual Φ é a distribuição acumulada da normal padrão, os parâmetros a e c determinam a curva ROC e r é um ponto de corte particular da curva ROC.

Demonstração: Consideramos que S_D representa os valores da variável de decisão para os indivíduos doentes e $S_{\overline{D}}$ representa os valores da variável de decisão para os indivíduos não doentes, em um teste diagnóstico cuja variável de decisão é S. Hipoteticamente, podemos descrever essa situação através da Figura A.2.

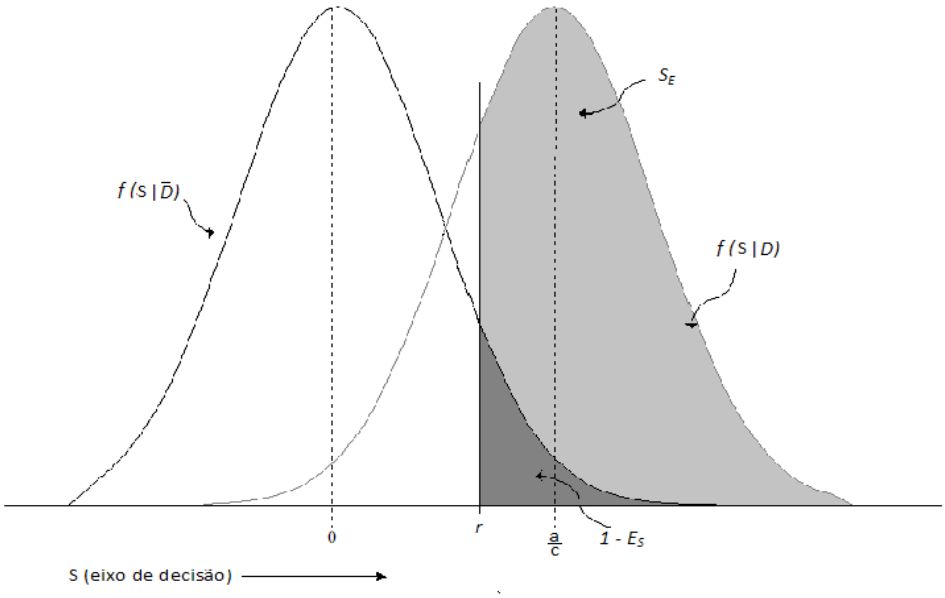

FIGURA A.2: Funções de densidade de probabilidade Gaussianas para os indivíduos doentes e para os indivíduos não doentes.

$$(1 - E_S)(r) = \int_r^{+\infty} f(s \mid \overline{D})ds = \Phi(+\infty) - \Phi(z_{\overline{D}})_r$$
$$= 1 - \Phi\left(\frac{r - \mu_{\overline{D}}}{\sigma_{\overline{D}}}\right). \tag{A.2}$$

Como vimos anteriormente, temos que $S_{\overline{D}} \sim N(0\ ,\ 1)$. Logo, da equação A.2, podemos concluir que

$$\mu_{\overline{D}} = 0 \quad e \quad \sigma_{\overline{D}} = 1.$$

Assim,

$$(1 - E_S)(r) = 1 - \Phi\left(\frac{r - \mu_{\overline{D}}}{\sigma_{\overline{D}}}\right)$$
$$= 1 - \Phi\left(\frac{r - 0}{1}\right) = 1 - \Phi(r) = \Phi(-r).$$

$$(S_E)(r) = \int_r^{+\infty} f(s \mid D)ds = \Phi(+\infty) - \Phi(z_D)_r$$
$$= 1 - \Phi\left(\frac{r - \mu_D}{\sigma_D}\right). \tag{A.3}$$

Como vimos anteriormente, temos que $S_D \sim N\left(\frac{a}{c}\ ,\ \frac{1}{c}\right)$. Logo, da equação A.3, podemos concluir que

$$\mu_D = \frac{a}{c} \quad e \quad \sigma_D = \frac{1}{c}.$$

Assim,

$$(S_E)(r) = 1 - \Phi\left(\frac{r - \mu_D}{\sigma_D}\right) = 1 - \Phi\left(\frac{r - \frac{a}{c}}{\frac{1}{c}}\right)$$
$$= 1 - \Phi(cr - a) = \Phi[-(cr - a)] = \Phi(a - cr). \qquad \blacksquare$$

Apêndice B

Teste de *Kolmogorov-Smirnov*

Segundo Conover (1971), o teste de *Kolmogorov-Smirnov* observa a diferença absoluta máxima entre a função de distribuição acumulada, assumida para os dados, e a função de distribuição empírica dos dados. Como critério, comparamos esta diferença com um valor crítico (Tabela B.1), para um dado nível de significância.

A estatística utilizada para o teste é

$$Dist_q \;=\; \sup_h \mid F(h) - F_q(h) \mid,$$

em que $Dist_q$ corresponde à distância máxima vertical entre os gráficos de $F(h)$ e $F_q(h)$, sobre a amplitude dos possíveis valores de h, $F(h)$ representa a função de distribuição acumulada assumida para os dados e $F_q(h)$ representa a função de distribuição acumulada empírica dos dados.

Sejam $H_{(1)} \leq H_{(2)} \leq \cdots \leq H_{(q)}$ observações aleatórias ordenadas de forma crescente da variável aleatória contínua H. A função de distribuição acumulada assumida para os dados é definida por $F(h_{(q^{\blacktriangle})}) = P(H \leq h_{(q^{\blacktriangle})})$ e a função de distribuição acumulada empírica é definida por uma função escada

$$F_q(h) \;=\; \frac{1}{q} \sum_{q^{\blacktriangle}=1}^{q} I_{\{-\infty,\, h\}}\, (h_{(q^{\blacktriangle})}), \tag{B.1}$$

em que $I_{\{A\}}$ é a função indicadora. A função indicadora é definida da seguinte

forma

$$I_{\{A\}}(h) = \begin{cases} 1, & \text{se } h \in A, \\ 0, & \text{caso contrário.} \end{cases}$$

Observamos que a função da distribuição empírica $F_q(h)$ corresponde à proporção de valores menores ou iguais a h.

A expressão B.1 pode também ser escrita por

$$F_q(h) = \begin{cases} 0, & \text{se } h < h_{(1)}, \\ \dfrac{k}{q}, & \text{se } h_{(k)} \leq h < h_{(k+1)}, \\ 1, & \text{se } h > h_{(q)}. \end{cases}$$

Sejam

$$Dist^{+} = \sup_{h_{(q^{\blacktriangle})}} \mid F(h_{(q^{\blacktriangle})}) - F_q(h_{(q^{\blacktriangle})}) \mid$$

e

$$Dist^{-} = \sup_{h_{(q^{\blacktriangle})}} \mid F(h_{(q^{\blacktriangle})}) - F_q(h_{(q^{\blacktriangle}-1)}) \mid$$

duas estatísticas que medem as distâncias verticais entre o gráfico da função teórica e o gráfico da função empírica, nos pontos $h_{(q^{\blacktriangle}-1)}$ e $h_{(q^{\blacktriangle})}$. Podemos utilizar $Dist_q = \max\left(Dist^{+}, Dist^{-}\right)$ como uma outra estatística do teste de *Kolmogorov-Smirnov*.

Uma regra de decisão baseada nessa nova estatística consiste em: se $Dist_q$ for maior que o valor crítico encontrado na Tabela B.1, rejeitamos a hipótese de normalidade dos dados com $(1 - \alpha)$ de confiança, caso contrário, não rejeitamos a hipótese de normalidade.

TABELA B.1: Valores críticos para a estatística $Dist_q$ do teste de *Kolmogorov-Smirnov*

	Nível de significância (α)			
q	*0,2*	*0,1*	*0,05*	*0,01*
5	0,45	0,51	0,56	0,67
10	0,32	0,37	0,41	0,49
15	0,27	0,30	0,34	0,40
20	0,23	0,26	0,29	0,36
25	0,21	0,24	0,27	0,32
30	0,19	0,22	0,24	0,29
35	0,18	0,20	0,23	0,27
40	0,17	0,19	0,21	0,25
45	0,16	0,18	0,20	0,24
50	0,15	0,17	0,19	0,23
> 50	$\frac{1,07}{\sqrt{n}}$	$\frac{1,22}{\sqrt{n}}$	$\frac{1,36}{\sqrt{n}}$	$\frac{1,63}{\sqrt{n}}$

Podemos reescrever os dados da Tabela B.1 referentes à estatística $Dist_q$. As novas estatísticas $Dist_q$ encontram-se na Tabela B.2.

TABELA B.2: Resumo do cálculo de $Dist_q$

h (ordenado)	$F_q(h)$	$F(h) = P\left(z_{(q^\blacktriangle)} \leq \frac{h_{(q^\blacktriangle)} - \overline{h}}{desvio}\right)$	$\lvert F(h_{(q^\blacktriangle)}) - F_q(h_{(q^\blacktriangle)})\rvert$	$\lvert F(h_{(q^\blacktriangle)}) - F_q(h_{(q^\blacktriangle - 1)})\rvert$
$h_{(1)}$	$\frac{1}{q}$	$F(h) = P\left(z_{(1)} \leq \frac{h_{(1)} - \overline{h}}{desvio}\right)$	$\lvert F(h_{(1)}) - F_q(h_{(1)})\rvert$	$\lvert F(h_{(1)}) - 0)\rvert$
$h_{(2)}$	$\frac{2}{q}$	$F(h) = P\left(z_{(2)} \leq \frac{h_{(2)} - \overline{h}}{desvio}\right)$	$\lvert F(h_{(2)}) - F_q(h_{(2)})\rvert$	$\lvert F(h_{(2)}) - F_q(h_{(1)})\rvert$
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
$h_{(q-1)}$	$\frac{q-1}{q}$	$F(h) = P\left(z_{(q-1)} \leq \frac{h_{(q-1)} - \overline{h}}{desvio}\right)$	$\lvert F(h_{(q-1)}) - F_q(z_{(q-1)})\rvert$	$\lvert F(h_{(q-1)}) - F_q(h_{(q-2)})\rvert$
$h_{(q)}$	1	$F(h) = P\left(z_{(q)} \leq \frac{h_{(q)} - \overline{h}}{desvio}\right)$	$\lvert F(h_{(q)}) - F_q(h_{(q)})\rvert$	$\lvert F(h_{(q)}) - F_q(h_{(q-1)})\rvert$

Observação: O valor de $P\left(Z_{(q^\blacktriangle)} \leq \frac{h_{(q^\blacktriangle)} - \overline{h}}{desvio}\right)$ é encontrado na tabela da distribuição normal padrão.

Apêndice C

Programas

C.1 Rotinas referentes à Seção 6.2.3

C.1.1 Método das médias verticais (VA)

```
# LEITURA DOS DADOS
dados  <- read.table("dados.txt", header=T)
status <- dados$d
teste  <- dados$t

# FUNÇÃO QUE CALCULA A TAXA DE FP E VP PARA UM DADO PTO DE CORTE
vp_fp <- function(status, teste, pto) {
    clasf <- factor(x = as.numeric(teste > pto), levels = c(0, 1))
    tab <- table(status, clasf);
    vp <- tab["1","1"]/sum(tab["1",])
    fp <- tab["0","1"]/sum(tab["0",])
    return(list(vp = vp, fp = fp))
}

# FUNÇÃO QUE RETORNA O GRÁFICO DA CURVA ROC PARA UM DADO VALOR DE FP
graf_roc <- function(status, teste, cortes = 0:5, primeiro = TRUE){
   POS <- matrix(0, nrow = length(cortes), ncol = 2)
   for (i in 1:length(cortes)) {
       obj <- vp_fp(status, teste, cortes[i])
       POS[i, ] <- c(obj$fp, obj$vp)
   }
   if (primeiro)
     plot(POS[,1], POS[,2], main = "(a)", type = "l", xlab = "FP", ylab = "VP", col = "black", lwd = 2) else
     lines(POS[,1], POS[,2], col = gray(0.6))
   return(POS)
}
```

```
# FUNÇÃO QUE RETORNA O VALOR DE VP PARA UM DADO VALOR DE FP USANDO INTERPOLAÇÃO LINEAR
roc <- function(status, teste, fp, cortes = 0:5){
   POS <- matrix(0, nrow = length(cortes), ncol = 2)
   for (i in 1:length(cortes)) {
       obj <- vp_fp(status, teste, cortes[i])
       POS[i, ] <- c(obj$fp, obj$vp)
   }
   POS <- POS[!duplicated(POS), ]
   ind <- order(POS[,1], POS[,2])
   POS <- POS[ind, ]
   if (fp == 1)
      vp <- 1 else {
      ind_x <- max(which(POS[ ,1] <= fp))
      x1 <- POS[ind_x ,1]; x2 <-  POS[ind_x + 1, 1]
      y1 <- POS[ind_x ,2]; y2 <-  POS[ind_x + 1, 2]
      vp <- y1 + (fp - x1)*(y2 - y1)/(x2 - x1)
   }
   return(vp)
}

M = 10000 # NÚMERO DE AMOSTRAS BOOTSTRAP

dados_boot <- vector("list", length = M) # ARRAY DOS DADOS BOOTSTRAP
graf_roc(status, teste, primeiro = TRUE)
set.seed(270183)
for (i in 1:M){
    amostra = sample(nrow(dados), replace=T)
    dados_n = dados[amostra,]
    status_n <- dados_n[,1]
    teste_n  <- dados_n[, 2]
    dados_boot[[i]] <- dados_n
    graf_roc(status_n, teste_n, primeiro = FALSE)
}

delta <- 0.05 # NÍVEL DE SIGNIFICÂNCIA
x <- seq(0, 1, by = 0.01)
y <- numeric(length(x)) # VALORES MÉDIOS DE VP
li1 <- li2 <- li3 <- numeric(length(x)) # LIMITES INFERIORES PARA VP
ls1 <- ls2 <- ls3 <- numeric(length(x)) # LIMITES SUPERIORES PARA VP

for (i in 1:length(x)) {
    vals <- numeric()
    for (j in 1:M) {
         vals <- c(vals, roc(dados_boot[[j]]$d, dados_boot[[j]]$t, fp = x[i]))
    }
    vals <- sort(vals)
    y[i] <- mean(vals)
    sigma <- sd(vals)
```

```
    # DISTRIBUIÇÃO EMPÍRICA
    li1[i] <- vals[floor((delta/2)*M)]
    ls1[i] <- vals[ceiling((1 - delta/2)*M)]

    # DISTRIBUIÇÃO BINOMIAL
    nd <- sum(status == 1) #NÚMERO DE INDIVÍDUOS DOENTES
    z_delta <- qnorm(1 - delta/2)
    li2[i] <- y[i] - z_delta*sqrt(y[i]*(1 - y[i])/nd)
    ls2[i] <- y[i] + z_delta*sqrt(y[i]*(1 - y[i])/nd)

    # DISTRIBUIÇÃO NORMAL
    t_delta <- qt(1 - delta/2, nd - 1)
    li3[i] <- y[i] - t_delta*sigma
    ls3[i] <- y[i] + t_delta*sigma
}

# GRÁFICO DA CURVA ROC MÉDIA E DAS BANDAS DE CONFIANÇA VA
lines(x, y, lwd = 2, lty = 5) # CURVA ROC MÉDIA
# DISTRIBUIÇÃO EMPÍRICA
lines(x, li1, lwd = 2, lty = 6)
lines(x, ls1, lwd = 2, lty = 6)
# DISTRIBUIÇÃO BINOMIAL
lines(x, li2, lwd = 2, lty = 3)
lines(x, ls2, lwd = 2, lty = 3)
# DISTRIBUIÇÃO NORMAL
lines(x, li3, lwd = 2, lty = 4)
lines(x, ls3, lwd = 2, lty = 4)

# CÁLCULO DAS PROBABILIDADES DE COBERTURA
# VERIFICA SE TODOS OS PONTOS DA REAMOSTRA j ESTÃO CONTIDOS NA BANDA DE CONFIANÇA PARA CADA MÉTODO
cont1 <- cont2 <- cont3 <- 0
for (i in 1:M) {
     indicador1 <- indicador2 <- indicador3 <- 1
     for (j in 1:length(x)) {
         pto <- roc(dados_boot[[i]]$d, dados_boot[[i]]$t, fp = x[j])
         if ((pto < li1[j]) || (pto > ls1[j])) indicador1 <- 0
         if ((pto < li2[j]) || (pto > ls2[j])) indicador2 <- 0
         if ((pto < li3[j]) || (pto > ls3[j])) indicador3 <- 0
     }

     # ACUMULA O NÚMERO DE CURVAS ROC QUE ESTÃO DENTRO DA BANDA DE CONFIANÇA PARA CADA MÉTODO
     cont1 <- cont1 + indicador1
     cont2 <- cont2 + indicador2
     cont3 <- cont3 + indicador3
}

# VALOR DA PROBABILIDADE DE COBERTURA PARA CADA DISTRIBUIÇÃO USADA PELO MÉTODO VA
prob1 <- cont1/M
```

```
prob2 <- cont2/M
prob3 <- cont3/M
```

C.1.2 Método das médias otimizadas (MO)

```
# LEITURA DOS DADOS
dados  <- read.table("dados.txt", header=T)
status <- dados$d
teste  <- dados$t

# FUNÇÃO QUE CALCULA A TAXA DE FP E VP PARA UM DADO PTO DE CORTE
vp_fp <- function(status, teste, pto) {
    clasf <- factor(x = as.numeric(teste > pto), levels = c(0, 1))
    tab <- table(status, clasf);
    vp <- tab["1","1"]/sum(tab["1",])
    fp <- tab["0","1"]/sum(tab["0",])
    return(list(vp = vp, fp = fp))
}

# FUNÇÃO QUE RETORNA O GRÁFICO DA CURVA ROC PARA UM DADO VALOR DE FP
graf_roc <- function(status, teste, cortes = 0:5, primeiro = TRUE){
   POS <- matrix(0, nrow = length(cortes), ncol = 2)
   for (i in 1:length(cortes)) {
       obj <- vp_fp(status, teste, cortes[i])
       POS[i, ] <- c(obj$fp, obj$vp)
   }
   if (primeiro)
     plot(POS[,1], POS[,2], main = "(a)", type = "l", xlab = "FP", ylab = "VP", col = "black", lwd = 2) else
     lines(POS[,1], POS[,2], col = gray(0.6))
   return(POS)
}

# FUNÇÃO QUE RETORNA O VALOR DE VP PARA UM DADO VALOR DE FP USANDO INTERPOLAÇÃO LINEAR
roc_vp <- function(status, teste, fp, cortes = 0:5){
   POS <- matrix(0, nrow = length(cortes), ncol = 2)
   for (i in 1:length(cortes)) {
       obj <- vp_fp(status, teste, cortes[i])
       POS[i, ] <- c(obj$fp, obj$vp)
   }
   POS <- POS[!duplicated(POS), ]
   ind <- order(POS[,1], POS[,2])
   POS <- POS[ind, ]
   if (fp == 1)
      vp <- 1 else {
      ind_fp <- max(which(POS[ ,1] <= fp))
      x1 <- POS[ind_fp ,1]; x2 <-  POS[ind_fp + 1, 1]
      y1 <- POS[ind_fp ,2]; y2 <-  POS[ind_fp + 1, 2]
      vp <- y1 + (fp - x1)*(y2 - y1)/(x2 - x1)
```

```
   }
   return(vp)
}

# FUNÇÃO QUE RETORNA O VALOR DE FP PARA UM DADO VALOR DE VP USANDO INTERPOLAÇÃO LINEAR
roc_fp <- function(status, teste, vp, cortes = 0:5){
   POS <- matrix(0, nrow = length(cortes), ncol = 2)
   for (i in 1:length(cortes)) {
       obj <- vp_fp(status, teste, cortes[i])
       POS[i, ] <- c(obj$fp, obj$vp)
   }
   POS <- POS[!duplicated(POS), ]
   ind <- order(POS[,1], POS[,2])
   POS <- POS[ind, ]
   if (vp == 1)
      fp <- 1 else {
      ind_vp <- max(which(POS[ ,2] <= vp))
      x1 <- POS[ind_vp ,1]; x2 <-  POS[ind_vp + 1, 1]
      y1 <- POS[ind_vp ,2]; y2 <-  POS[ind_vp + 1, 2]
      fp <- x1 + (vp - y1)*(x2 - x1)/(y2 - y1)
   }
   return(fp)
}

interpola <- function(Mat, x) {
   Mat <- Mat[!duplicated(Mat), ]
   ind <- order(Mat[,1], Mat[,2])
   Mat <- Mat[ind, ]

   # CALCULA O VALOR DE y VIA INTERPOLAÇÃO LINEAR DOS PONTOS EM Mat
   if (any(which(Mat[ ,1] <= x))) {
      ind_x <- max(which(Mat[ ,1] <= x))
      if (x == Mat[ind_x, 1]) {
         y <- Mat[ind_x, 2]
      } else {
         if (ind_x == nrow(Mat)) {
            x1 <- Mat[ind_x - 1, 1]; x2 <- Mat[ind_x, 1]
            y1 <- Mat[ind_x - 1, 2]; y2 <- Mat[ind_x, 2]
            if (x2 == x1) {
                y <- y2
            } else {
                y  <- y1 + (x - x1)*(y2 - y1)/(x2 - x1)
            }
         } else {
            x1 <- Mat[ind_x, 1]; x2 <- Mat[ind_x + 1, 1]
            y1 <- Mat[ind_x, 2]; y2 <- Mat[ind_x + 1, 2]
            y  <- y1 + (x - x1)*(y2 - y1)/(x2 - x1)
         }
      }
```

```
   } else {
      x1 <- Mat[1, 1]; x2 <- Mat[2, 1]
      y1 <- Mat[1, 2]; y2 <- Mat[2, 2]
      if (x2 == x1) {
          y <- y1
      } else {
          y  <- y1 + (x - x1)*(y2 - y1)/(x2 - x1)
      }
   }
   return(y)
}

M = 10000 # NÚMERO DE AMOSTRAS BOOTSTRAP

dados_boot <- vector("list", length = M) # ARRAY DOS DADOS BOOTSTRAP
graf_roc(status, teste, primeiro = TRUE)
set.seed(270183)
for (i in 1:M){
    amostra = sample(nrow(dados), replace=T)
    dados_n = dados[amostra,]
    status_n <- dados_n[,1]
    teste_n  <- dados_n[, 2]
    dados_boot[[i]] <- dados_n
    graf_roc(status_n, teste_n, primeiro = FALSE)
}

delta <- 0.05 # NÍVEL DE SIGNIFICÂNCIA
x <- seq(0, 1, by = 0.01)
y <- numeric(length(x)) # VALORES MÉDIOS DE VP
x_barra <- numeric(length(x))
li1_x <- li2_x <- li3_x <- numeric(length(x)) # LIMITES INFERIORES PARA FP
li1_y <- li2_y <- li3_y <- numeric(length(x)) # LIMITES INFERIORES PARA VP
ls1_x <- ls2_x <- ls3_x <- numeric(length(x)) # LIMITES SUPERIORES PARA FP
ls1_y <- ls2_y <- ls3_y <- numeric(length(x)) # LIMITES SUPERIORES PARA VP

for (i in 1:length(x)) {
    vals_y <- numeric()
    for (j in 1:M) {
         vals_y <- c(vals_y, roc_vp(dados_boot[[j]]$d, dados_boot[[j]]$t, fp = x[i]))
    }
    vals_y <- sort(vals_y)
    y[i] <- mean(vals_y)
    sigma_y <- sd(vals_y)
    vals_x <- numeric()
    for (j in 1:M) {
         vals_x <- c(vals_x, roc_fp(dados_boot[[j]]$d, dados_boot[[j]]$t, vp = y[i]))
    }
    vals_x <- sort(vals_x)
    x_barra[i] <- mean(vals_x)
```

```
    sigma_x <- sd(vals_x)

    # DISTRIBUIÇÃO EMPÍRICA
    li1_y[i] <- vals_y[floor((delta/2)*M)]
    ls1_y[i] <- vals_y[ceiling((1 - delta/2)*M)]
    li1_x[i] <- vals_x[floor((delta/2)*M)]
    ls1_x[i] <- vals_x[ceiling((1 - delta/2)*M)]

    # DISTRIBUIÇÃO BINOMIAL
    nd <- sum(status == 1)   # NÚMERO DE INDIVÍDUOS DOENTES
    n_nd <- sum(status == 0) # NÚMERO DE INDIVÍDUOS NÃO DOENTES
    z_delta <- qnorm(1 - delta/2)
    li2_y[i] <- y[i] - z_delta*sqrt(y[i]*(1 - y[i])/nd)
    ls2_y[i] <- y[i] + z_delta*sqrt(y[i]*(1 - y[i])/nd)
    li2_x[i] <- x_barra[i] - z_delta*sqrt(x[i]*(1 - x[i])/n_nd)
    ls2_x[i] <- x_barra[i] + z_delta*sqrt(x[i]*(1 - x[i])/n_nd)

    # DISTRIBUIÇÃO NORMAL
    ty_delta <- qt(1 - delta/2, nd - 1)
    tx_delta <- qt(1 - delta/2, n_nd - 1)
    li3_y[i] <- y[i] - ty_delta*sigma_y
    ls3_y[i] <- y[i] + ty_delta*sigma_y
    li3_x[i] <- x_barra[i] - tx_delta*sigma_x
    ls3_x[i] <- x_barra[i] + tx_delta*sigma_x
}

# MATRIZES PARA DETERMINAÇÃO DA BANDA DE CONFIANÇA
M1_sup <- cbind((x + li1_x)/2, (y + ls1_y)/2)
M1_inf <- cbind((x + ls1_x)/2, (y + li1_y)/2)
M2_sup <- cbind((x + li2_x)/2, (y + ls2_y)/2)
M2_inf <- cbind((x + ls2_x)/2, (y + li2_y)/2)
M3_sup <- cbind((x + li3_x)/2, (y + ls3_y)/2)
M3_inf <- cbind((x + ls3_x)/2, (y + li3_y)/2)

# ORDENAÇÃO DAS MATRIZES
M1_sup <- M1_sup[!duplicated(M1_sup), ]
ind <- order(M1_sup[,1], M1_sup[,2])
M1_sup <- M1_sup[ind, ]
M1_inf <- M1_inf[!duplicated(M1_inf), ]
ind <- order(M1_inf[,1], M1_inf[,2])
M1_inf <- M1_inf[ind, ]

M2_sup <- M2_sup[!duplicated(M2_sup), ]
ind <- order(M2_sup[,1], M2_sup[,2])
M2_sup <- M2_sup[ind, ]
M2_inf <- M2_inf[!duplicated(M2_inf), ]
ind <- order(M2_inf[,1], M2_inf[,2])
M2_inf <- M2_inf[ind, ]
```

```
M3_sup <- M3_sup[!duplicated(M3_sup), ]
ind <- order(M3_sup[,1], M3_sup[,2])
M3_sup <- M3_sup[ind, ]
M3_inf <- M3_inf[!duplicated(M3_inf), ]
ind <- order(M3_inf[,1], M3_inf[,2])
M3_inf <- M3_inf[ind, ]

# GRÁFICO DAS BANDAS DE CONFIANÇA MO
# DISTRIBUIÇÃO EMPÍRICA
lines(M1_sup[,1], M1_sup[,2], lwd = 2, lty = 6)
lines(M1_inf[,1], M1_inf[,2], lwd = 2, lty = 6)
# DISTRIBUIÇÃO BINOMIAL
lines(M2_sup[,1], M2_sup[,2], lwd = 2, lty = 3)
lines(M2_inf[,1], M2_inf[,2], lwd = 2, lty = 3)
# DISTRIBUIÇÃO NORMAL
lines(M3_sup[,1], M3_sup[,2], lwd = 2, lty = 4)
lines(M3_inf[,1], M3_inf[,2], lwd = 2, lty = 4)

# CÁLCULO DAS PROBABILIDADES DE COBERTURA
cortes = 0:5
cont1 <- cont2 <- cont3 <- 0
for (j in 1:M) {
    # CALCULA A MATRIZ DOS PONTOS DA CURVA ROC PARA REAMOSTRA j
    obj <- dados_boot[[j]]
    status_n <- obj$d
    teste_n <-  obj$t
    Matriz <- matrix(0, nrow = 6, ncol = 2)
    for (i in 1:length(cortes)) {
        obj <- vp_fp(status_n, teste_n, pto = cortes[i])
        Matriz[i, ] <- c(obj$fp, obj$vp)
    }

    # REMOVE AS LINHAS REPETIDAS (SE HOUVER) E ORDENA AS LINHAS RESTANTES
    Matriz <- Matriz[!duplicated(Matriz), ]
    ind <- order(Matriz[,1], Matriz[,2])
    Matriz <- Matriz[ind, ]
    x_Matriz <- unique(Matriz[, 1])

    # VERIFICA SE TODOS OS PONTOS DA REAMOSTRA j ESTÃO CONTIDOS NA BANDA DE CONFIANÇA PARA CADA MÉTODO
    indicador1 <- 1
    indicador2 <- 1
    indicador3 <- 1
    for (i in 1:(length(x_Matriz))) {
         if (x_Matriz[i] > 0) {
             y_Matriz <- roc_vp(status_n, teste_n, fp = x_Matriz[i])

             y1_inf <- interpola(M1_inf, x_Matriz[i])
             y1_sup <- interpola(M1_sup, x_Matriz[i])
             if ((y_Matriz > y1_sup) || (y_Matriz < y1_inf)) indicador1 <- 0
```

```
            y2_inf <- interpola(M2_inf, x_Matriz[i])
            y2_sup <- interpola(M2_sup, x_Matriz[i])
            if ((y_Matriz > y2_sup) || (y_Matriz < y2_inf)) indicador2 <- 0

            y3_inf <- interpola(M3_inf, x_Matriz[i])
            y3_sup <- interpola(M3_sup, x_Matriz[i])
            if ((y_Matriz > y3_sup) || (y_Matriz < y3_inf)) indicador3 <- 0
        }
    }

    # ACUMULA O NÚMERO DE CURVAS ROC QUE ESTÃO DENTRO DA BANDA DE CONFIANÇA PARA CADA MÉTODO
    cont1 <- cont1 + indicador1
    cont2 <- cont2 + indicador2
    cont3 <- cont3 + indicador3
}

# VALOR DA PROBABILIDADE DE COBERTURA PARA CADA DISTRIBUIÇÃO USADA PELO MÉTODO MO
prob1 <- cont1/M
prob2 <- cont2/M
prob3 <- cont3/M
```

C.1.3 Método das médias limiares (TA)

```
# LEITURA DOS DADOS
dados  <- read.table("dados.txt", header=T)
status <- dados$d
teste  <- dados$t

# FUNÇÃO QUE CALCULA A TAXA DE FP E VP PARA UM DADO PTO DE CORTE
vp_fp <- function(status, teste, pto) {
    clasf <- factor(x = as.numeric(teste > pto), levels = c(0, 1))
    tab <- table(status, clasf);
    vp <- tab["1","1"]/sum(tab["1",])
    fp <- tab["0","1"]/sum(tab["0",])
    return(list(vp = vp, fp = fp))
}

# FUNÇÃO QUE RETORNA O GRÁFICO DA CURVA ROC PARA UM DADO VALOR DE FP
graf_roc <- function(status, teste, cortes = 0:5, original = TRUE){
   POS <- matrix(0, nrow = length(cortes), ncol = 2)
   for (i in 1:length(cortes)) {
       obj <- vp_fp(status, teste, cortes[i])
       POS[i, ] <- c(obj$fp, obj$vp)
   }
   if (original) cor = "black" else cor = gray(0.6)
   lines(POS[,1], POS[,2], col = gray(0.6), lwd = 1)
}
```

```
# FUNÇÃO QUE RETORNA O VALOR DE VP PARA UM DADO VALOR DE FP
roc <- function(POS, fp){
   POS <- POS[!duplicated(POS), ]
   ind <- order(POS[,1], POS[,2])
   POS <- POS[ind, ]
   if (fp == 1)
      vp <- 1 else {
      ind_x <- max(which(POS[ ,1] <= fp))
      x1 <- POS[ind_x ,1]; x2 <-  POS[ind_x + 1, 1]
      y1 <- POS[ind_x ,2]; y2 <-  POS[ind_x + 1, 2]
      vp <- y1 + (fp - x1)*(y2 - y1)/(x2 - x1)
   }
   return(vp)
}

M = 10000 # NÚMERO DE AMOSTRAS BOOTSTRAP

dados_boot <- vector("list", length = M) # ARRAY DOS DADOS BOOTSTRAP
set.seed(270183)
for (j in 1:M) {
    amostra <- sample(nrow(dados), replace=T)
    dados_n <- dados[amostra,]
    status_n <- dados_n[, 1]
    teste_n  <- dados_n[, 2]
    dados_boot[[j]] <- list(status = status_n, teste = teste_n)
}

# MATRIZ DOS TP's e VP's AMOSTRADOS VIA BOOTSTRAP
VP <- matrix(0, nrow = 6, ncol = M)
FP <- matrix(0, nrow = 6, ncol = M)
cortes <- 0:5
for (i in 1:6) {
    for (j in 1:M) {
        obj_boot <- dados_boot[[j]]
        status_n <- obj_boot$status
        teste_n  <- obj_boot$teste

        # VALOR DE VP PARA A AMOSTRA BOOTSTRAP j E PTO DE CORTE i
        obj <- vp_fp(status_n, teste_n, pto = cortes[i])
        VP[i, j] <- obj$vp
        FP[i, j] <- obj$fp
    }
}

delta <- 0.05 # NÍVEL DE SIGNIFICÂNCIA
x <- y <- numeric(6)
li1_vp <- li2_vp <- li3_vp <- numeric(6) # LIMITE INFERIOR PARA VP
ls1_vp <- ls2_vp <- ls3_vp <- numeric(6) # LIMITE SUPERIOR PARA VP
```

```
for (i in 1:6) {
    vals_vp <- sort(VP[i, ])
    vals_fp <- sort(FP[i, ])
    x[i] <- mean(vals_fp)
    y[i] <- mean(vals_vp)

    # DISTRIBUIÇÃO EMPÍRICA
    li1_vp[i] <- vals_vp[floor((delta/2)*M)]
    ls1_vp[i] <- vals_vp[ceiling((1 - delta/2)*M)]

    # DISTRIBUIÇÃO BINOMIAL
    nd <- sum(status == 1)  # NÚMERO DE INDIVÍDUOS DOENTES
    z_delta <- qnorm(1 - delta/2)
    li2_vp[i] <- y[i] - z_delta*sqrt(y[i]*(1 - y[i])/nd)
    ls2_vp[i] <- y[i] + z_delta*sqrt(y[i]*(1 - y[i])/nd)

    # DISTRIBUIÇÃO NORMAL
    sigma_vp <- sd(vals_vp)
    t_delta <- qt(1 - delta/2, nd - 1)
    li3_vp[i] <- y[i] - t_delta*sigma_vp
    ls3_vp[i] <- y[i] + t_delta*sigma_vp
}

# MATRIZES DOS PONTOS DAS CURVAS INFERIOR E SUPERIOR POR MÉTODO
POS1_inf <- cbind(x, li1_vp)
POS1_sup <- cbind(x, ls1_vp)
POS2_inf <- cbind(x, li2_vp)
POS2_sup <- cbind(x, ls2_vp)
POS3_inf <- cbind(x, li3_vp)
POS3_sup <- cbind(x, ls3_vp)

# CÁLCULO DAS PROBABILIDADES DE COBERTURA
prob1 <- prob2 <- prob3 <- 0
cont1 <- cont2 <- cont3 <- 0
for (j in 1:M) {
    # CALCULA A MATRIZ DOS PONTOS DA CURVA ROC PARA REAMOSTRA j
    obj <- dados_boot[[j]]
    status_n <- obj$status
    teste_n <-  obj$teste
    Matriz <- matrix(0, nrow = 6, ncol = 2)
    for (i in 1:6) {
        obj <- vp_fp(status_n, teste_n, pto = cortes[i])
        Matriz[i, ] <- c(obj$fp, obj$vp)
    }

    # OBTÉM OS VALORES DISTINTOS DE FP EM Matriz
    vals_x <- Matriz[, 1]
    vals_x <- vals_x[!duplicated(vals_x)]
```

```
    # CALCULA OS VALORES DE VP PARA OS VALORES vals_x
    vals_y <- sapply(vals_x, function(z) roc(Matriz, z))

    # VERIFICA SE TODOS OS PONTOS DA REAMOSTRA j ESTÃO CONTIDOS NA BANDA DE CONFIANÇA PARA CADA MÉTODO
    indicador1 <- indicador2 <- indicador3 <- 1
    for (i in 1:(length(vals_x))) {
        y1_inf <- roc(POS1_inf, vals_x[i])
        y1_sup <- roc(POS1_sup, vals_x[i])
        y2_inf <- roc(POS2_inf, vals_x[i])
        y2_sup <- roc(POS2_sup, vals_x[i])
        y3_inf <- roc(POS3_inf, vals_x[i])
        y3_sup <- roc(POS3_sup, vals_x[i])

        if ((vals_y[i] > y1_sup) || (vals_y[i] < y1_inf)) {indicador1 <- 0}
        if ((vals_y[i] > y2_sup) || (vals_y[i] < y2_inf)) {indicador2 <- 0}
        if ((vals_y[i] > y3_sup) || (vals_y[i] < y3_inf)) {indicador3 <- 0}
    }

    # ACUMULA O NÚMERO DE CURVAS ROC QUE ESTÃO DENTRO DA BANDA DE CONFIANÇA PARA CADA MÉTODO
    cont1 <- cont1 + indicador1
    cont2 <- cont2 + indicador2
    cont3 <- cont3 + indicador3
}

# VALOR DA PROBABILIDADE DE COBERTURA PARA CADA DISTRIBUIÇÃO USADA PELO MÉTODO TA
prob1 <- cont1/M
prob2 <- cont2/M
prob3 <- cont3/M

# GRÁFICO DA CURVA ROC ORIGINAL
plot(c(0,0), c(1,1), main = "(a)", type = "l", xlab = "FP", ylab = "VP", xlim = c(0,1), ylim = c(0,1))
for (i in 1:length(dados_boot)) {
    graf_roc(dados_boot[[i]]$status, dados_boot[[i]]$teste, original = FALSE)
}

# GRÁFICO DA CURVA ROC MÉDIA E BANDAS DE CONFIANÇA TA
graf_roc(status, teste)
lines(x, y, lwd = 2, lty = 5) # CURVA ROC MÉDIA
# DISTRIBUIÇÃO EMPÍRICA
lines(x, li1_vp, lwd = 2, lty = 6)
lines(x, ls1_vp, lwd = 2, lty = 6)
# DISTRIBUIÇÃO BINOMIAL
lines(x, li2_vp, lwd = 2, lty = 3)
lines(x, ls2_vp, lwd = 2, lty = 3)
# DISTRIBUIÇÃO NORMAL
lines(x, li3_vp, lwd = 2, lty = 4)
lines(x, ls3_vp, lwd = 2, lty = 4)
```

C.1.4 Método das médias limiares otimizadas (MLO)

```
# LEITURA DOS DADOS
dados  <- read.table("dados.txt", header=T)
status <- dados$d
teste  <- dados$t

# FUNÇÃO QUE CALCULA A TAXA DE FP E VP PARA UM DADO PTO DE CORTE
vp_fp <- function(status, teste, pto) {
    clasf <- factor(x = as.numeric(teste > pto), levels = c(0, 1))
    tab <- table(status, clasf);
    vp <- tab["1","1"]/sum(tab["1",])
    fp <- tab["0","1"]/sum(tab["0",])
    return(list(vp = vp, fp = fp))
}

# FUNÇÃO QUE RETORNA O GRÁFICO DA CURVA ROC PARA UM DADO VALOR DE FP
graf_roc <- function(status, teste, cortes = 0:5, original = TRUE){
   POS <- matrix(0, nrow = length(cortes), ncol = 2)
   for (i in 1:length(cortes)) {
       obj <- vp_fp(status, teste, cortes[i])
       POS[i, ] <- c(obj$fp, obj$vp)
   }
   if (original) cor = "black" else cor = gray(0.6)
   lines(POS[,1], POS[,2], col = cor, lwd = 1)
}

# FUNÇÃO QUE RETORNA O VALOR DE VP PARA UM DADO VALOR DE FP
roc <- function(POS, fp){
   POS <- POS[!duplicated(POS), ]
   ind <- order(POS[,1], POS[,2])
   POS <- POS[ind, ]
   if (fp == 1)
      vp <- 1 else {
      ind_x <- max(which(POS[ ,1] <= fp))
      x1 <- POS[ind_x ,1]; x2 <-  POS[ind_x + 1, 1]
      y1 <- POS[ind_x ,2]; y2 <-  POS[ind_x + 1, 2]
      vp <- y1 + (fp - x1)*(y2 - y1)/(x2 - x1)
   }
   return(vp)
}

M = 10000  # NÚMERO DE AMOSTRAS BOOTSTRAP

dados_boot <- vector("list", length = M) # ARRAY DOS DADOS BOOTSTRAP
set.seed(270183)
for (j in 1:M) {
    amostra <- sample(nrow(dados), replace=T)
    dados_n <- dados[amostra,]
```

```
    status_n <- dados_n[, 1]
    teste_n  <- dados_n[, 2]
    dados_boot[[j]] <- list(status = status_n, teste = teste_n)
}

# MATRIZ DOS TP's e VP's AMOSTRADOS VIA BOOTSTRAP
VP <- matrix(0, nrow = 6, ncol = M)
FP <- matrix(0, nrow = 6, ncol = M)
cortes <- 0:5
for (i in 1:6) {
    for (j in 1:M) {
        obj_boot <- dados_boot[[j]]
        status_n <- obj_boot$status
        teste_n  <- obj_boot$teste

        # VALOR DE VP PARA A AMOSTRA BOOTSTRAP j E PTO DE CORTE i
        obj <- vp_fp(status_n, teste_n, pto = cortes[i])
        VP[i, j] <- obj$vp
        FP[i, j] <- obj$fp
    }
}

delta <- 0.05 # NÍVEL DE SIGNIFICÂNCIA
x <- y <- numeric(6)
li1_vp <- li2_vp <- li3_vp <- numeric(6) # LIMITES INFERIORES PARA VP
li1_fp <- li2_fp <- li3_fp <- numeric(6) # LIMITES INFERIORES PARA FP
ls1_vp <- ls2_vp <- ls3_vp <- numeric(6) # LIMITES SUPERIORES PARA VP
ls1_fp <- ls2_fp <- ls3_fp <- numeric(6) # LIMITES SUPERIORES PARA FP

for (i in 1:6) {
    vals_vp <- sort(VP[i, ])
    vals_fp <- sort(FP[i, ])
    x[i] <- mean(vals_fp)
    y[i] <- mean(vals_vp)

    # DISTRIBUIÇÃO EMPÍRICA
    li1_vp[i] <- vals_vp[floor((delta/2)*M)]
    ls1_vp[i] <- vals_vp[ceiling((1 - delta/2)*M)]
    li1_fp[i] <- vals_fp[floor((delta/2)*M)]
    ls1_fp[i] <- vals_fp[ceiling((1 - delta/2)*M)]

    # DISTRIBUIÇÃO BINOMIAL
    nd <- sum(status == 1)  # NÚMERO DE INDIVÍDUOS DOENTES
    n_nd <- sum(status == 0) # NÚMERO DE INDIVÍDUOS NÃO DOENTES
    z_delta <- qnorm(1 - delta/2)
    li2_vp[i] <- y[i] - z_delta*sqrt(y[i]*(1 - y[i])/nd)
    ls2_vp[i] <- y[i] + z_delta*sqrt(y[i]*(1 - y[i])/nd)
    li2_fp[i] <- x[i] - z_delta*sqrt(x[i]*(1 - x[i])/n_nd)
    ls2_fp[i] <- x[i] + z_delta*sqrt(x[i]*(1 - x[i])/n_nd)
```

```
    # DISTRIBUIÇÃO NORMAL
    sigma_vp <- sd(vals_vp)
    sigma_fp <- sd(vals_fp)
    t_delta <- qt(1 - delta/2, nd - 1)
    li3_vp[i] <- y[i] - t_delta*sigma_vp
    ls3_vp[i] <- y[i] + t_delta*sigma_vp
    li3_fp[i] <- x[i] - t_delta*sigma_fp
    ls3_fp[i] <- x[i] + t_delta*sigma_fp
}

# MATRIZES DOS PONTOS DAS CURVAS INFERIOR E SUPERIOR POR MÉTODO
POS1_inf <- cbind(ls1_fp, li1_vp)
POS1_sup <- cbind(li1_fp, ls1_vp)
POS2_inf <- cbind(ls2_fp, li2_vp)
POS2_sup <- cbind(li2_fp, ls2_vp)
POS3_inf <- cbind(ls3_fp, li3_vp)
POS3_sup <- cbind(li3_fp, ls3_vp)

# ORDENAÇÃO DAS MATRIZES
POS1_inf <- POS1_inf[!duplicated(POS1_inf), ]
ind <- order(POS1_inf[,1], POS1_inf[,2])
POS1_inf <- POS1_inf[ind, ]

POS1_sup <- POS1_sup[!duplicated(POS1_sup), ]
ind <- order(POS1_sup[,1], POS1_sup[,2])
POS1_sup <- POS1_sup[ind, ]

POS2_inf <- POS2_inf[!duplicated(POS2_inf), ]
ind <- order(POS2_inf[,1], POS2_inf[,2])
POS2_inf <- POS2_inf[ind, ]

POS2_sup <- POS2_sup[!duplicated(POS2_sup), ]
ind <- order(POS2_sup[,1], POS2_sup[,2])
POS2_sup <- POS2_sup[ind, ]

POS3_inf <- POS3_inf[!duplicated(POS3_inf), ]
ind <- order(POS3_inf[,1], POS3_inf[,2])
POS3_inf <- POS3_inf[ind, ]

POS3_sup <- POS3_sup[!duplicated(POS3_sup), ]
ind <- order(POS3_sup[,1], POS3_sup[,2])
POS3_sup <- POS3_sup[ind, ]

# CÁLCULO DAS PROBABILIDADES DE COBERTURA
prob1 <- prob2 <- prob3 <- 0
cont1 <- cont2 <- cont3 <- 0
for (j in 1:M) {
    # CALCULA A MATRIZ DOS PONTOS DA CURVA ROC PARA REAMOSTRA j
```

```
    obj <- dados_boot[[j]]
    status_n <- obj$status
    teste_n <-  obj$teste
    Matriz <- matrix(0, nrow = 6, ncol = 2)
    for (i in 1:6) {
        obj <- vp_fp(status_n, teste_n, pto = cortes[i])
        Matriz[i, ] <- c(obj$fp, obj$vp)
    }
    ind <- order(Matriz[,1], Matriz[,2])
    Matriz <- Matriz[ind, ]

    # OBTÉM OS VALORES DISTINTOS DE FP EM Matriz
    vals_x <- Matriz[, 1]
    vals_x <- vals_x[!duplicated(vals_x)]

    # CALCULA OS VALORES DE VP PARA OS VALORES vals_x
    vals_y <- sapply(vals_x, function(z) roc(Matriz, z))

    # VERIFICA SE TODOS OS PONTOS DA REAMOSTRA j ESTÃO CONTIDOS NA BANDA DE CONFIANÇA PARA CADA MÉTODO
    indicador1 <- indicador2 <- indicador3 <- 1
    for (i in 1:(length(vals_x))) {
        if (vals_x[i] > -1) {
            y1_inf <- roc(POS1_inf, vals_x[i])
            y1_sup <- roc(POS1_sup, vals_x[i])
            y2_inf <- roc(POS2_inf, vals_x[i])
            y2_sup <- roc(POS2_sup, vals_x[i])
            y3_inf <- roc(POS3_inf, vals_x[i])
            y3_sup <- roc(POS3_sup, vals_x[i])

            if ((vals_y[i] > y1_sup) || (vals_y[i] < y1_inf)) {indicador1 <- 0}
            if ((vals_y[i] > y2_sup) || (vals_y[i] < y2_inf)) {indicador2 <- 0}
            if ((vals_y[i] > y3_sup) || (vals_y[i] < y3_inf)) {indicador3 <- 0}
        }
    }

    # ACUMULA O NÚMERO DE CURVAS ROC QUE ESTÃO DENTRO DA BANDA DE CONFIANÇA PARA CADA MÉTODO
    cont1 <- cont1 + indicador1
    cont2 <- cont2 + indicador2
    cont3 <- cont3 + indicador3
}

# VALOR DA PROBABILIDADE DE COBERTURA PARA CADA DISTRIBUIÇÃO USADA PELO MÉTODO MLO
prob1 <- cont1/M
prob2 <- cont2/M
prob3 <- cont3/M

# GRÁFICO DA CURVA ROC ORIGINAL E DAS BANDAS DE CONFIANÇA MLO
plot(c(0,0), c(1,1), main = "(a)", type = "l", xlab = "FP", ylab = "VP", xlim = c(0,1), ylim = c(0,1))
for (i in 1:length(dados_boot)) {
```

```
    graf_roc(dados_boot[[i]]$status, dados_boot[[i]]$teste, original = FALSE)
}
graf_roc(status, teste)
# DISTRIBUIÇÃO  EMPÍRICA
lines(POS1_inf[,1], POS1_inf[,2], lwd = 2, lty = 6)
lines(POS1_sup[,1], POS1_sup[,2], lwd = 2, lty = 6)
# DISTRIBUIÇÃO  BINOMIAL
lines(POS2_inf[,1], POS2_inf[,2], lwd = 2, lty = 3)
lines(POS2_sup[,1], POS2_sup[,2], lwd = 2, lty = 3)
# DISTRIBUIÇÃO  NORMAL
lines(POS3_inf[,1], POS3_inf[,2], lwd = 2, lty = 4)
lines(POS3_sup[,1], POS3_sup[,2], lwd = 2, lty = 4)
```

C.1.5 Método das juntas simultâneas (SJR)

```
# LEITURA DOS DADOS
dados  <- read.table("dados.txt", header=T)
status <- dados$d
teste  <- dados$t

# FUNÇÃO QUE CALCULA A TAXA DE FP E VP PARA UM DADO PTO DE CORTE
vp_fp <- function(status, teste, pto) {
    clasf <- factor(x = as.numeric(teste > pto), levels = c(0, 1))
    tab <- table(status, clasf)
    vp <- tab["1","1"]/sum(tab["1",])
    fp <- tab["0","1"]/sum(tab["0",])
    return(list(vp = vp, fp = fp))
}

# FUNÇÃO QUE RETORNA O GRÁFICO DA CURVA ROC PARA UM DADO VALOR DE FP
graf_roc <- function(status, teste, cortes = 0:5, original = TRUE){
   Mat <- matrix(0, nrow = length(cortes), ncol = 2)
   for (i in 1:length(cortes)) {
       obj <- vp_fp(status, teste, cortes[i])
       Mat[i, ] <- c(obj$fp, obj$vp)
   }

   # CONSTRUÇÃO DA CURVA ROC
   if (original) cor = "black" else cor = gray(0.6)
   lines(Mat[,1], Mat[,2], col = gray(0.6), lwd = 1)
}

# FUNÇÃO QUE RETORNA O VALOR DE VP PARA UM DADO VALOR DE FP USANDO INTERPOLAÇÃO LINEAR
roc <- function(Mat, fp){
   Mat <- Mat[!duplicated(Mat), ]
   ind <- order(Mat[,1], Mat[,2])
   Mat <- Mat[ind, ]
   if (fp == 1)
```

```
      vp <- 1 else {
        ind_x <- max(which(Mat[ ,1] <= fp))
        x1 <- Mat[ind_x ,1]; x2 <- Mat[ind_x + 1, 1]
        y1 <- Mat[ind_x ,2]; y2 <- Mat[ind_x + 1, 2]
        vp <- y1 + (fp - x1)*(y2 - y1)/(x2 - x1)
   }
   return(vp)
}

interpola <- function(Mat, x) {
   Mat <- Mat[!duplicated(Mat), ]
   ind <- order(Mat[,1], Mat[,2])
   Mat <- Mat[ind, ]

   # CALCULA O VALOR DE y VIA INTERPOLAÇÃO LINEAR DOS PONTOS EM Mat
   if (any(which(Mat[ ,1] <= x))) {
      ind_x <- max(which(Mat[ ,1] <= x))
      if (x == Mat[ind_x, 1]) {
         y <- Mat[ind_x, 2]
      } else {
         if (ind_x == nrow(Mat)) {
            x1 <- Mat[ind_x - 1, 1]; x2 <- Mat[ind_x, 1]
            y1 <- Mat[ind_x - 1, 2]; y2 <- Mat[ind_x, 2]
            if (x2 == x1) {
                y <- y2
            } else {
                y  <- y1 + (x - x1)*(y2 - y1)/(x2 - x1)
            }
         } else {
            x1 <- Mat[ind_x, 1]; x2 <- Mat[ind_x + 1, 1]
            y1 <- Mat[ind_x, 2]; y2 <- Mat[ind_x + 1, 2]
            y  <- y1 + (x - x1)*(y2 - y1)/(x2 - x1)
         }
      }
   } else {
      x1 <- Mat[1, 1]; x2 <- Mat[2, 1]
      y1 <- Mat[1, 2]; y2 <- Mat[2, 2]
      if (x2 == x1) {
          y <- y1
      } else {
          y  <- y1 + (x - x1)*(y2 - y1)/(x2 - x1)
      }
   }
   return(y)
}

M = 10000 # NÚMERO DE AMOSTRAS BOOTSTRAP

dados_boot <- vector("list", length = M) # ARRAY DOS DADOS BOOTSTRAP
```

```
set.seed(270183)
for (j in 1:M) {
    amostra <- sample(nrow(dados), replace=T)
    dados_n <- dados[amostra,]
    status_n <- dados_n[, 1]
    teste_n  <- dados_n[, 2]
    dados_boot[[j]] <- list(status = status_n, teste = teste_n)
}

# MATRIZ DOS TP's e VP's AMOSTRADOS VIA BOOTSTRAP
VP <- matrix(0, nrow = 6, ncol = M)
FP <- matrix(0, nrow = 6, ncol = M)
cortes <- 0:5
for (i in 1:6) {
    for (j in 1:M) {
        obj_boot <- dados_boot[[j]]
        status_n <- obj_boot$status
        teste_n  <- obj_boot$teste

        # VALOR DE VP PARA A AMOSTRA BOOTSTRAP j E PTO DE CORTE i
        obj <- vp_fp(status_n, teste_n, pto = cortes[i])
        VP[i, j] <- obj$vp
        FP[i, j] <- obj$fp
    }
}

# CRIAÇÃO DA MATRIZ DE COORDENADAS DA CURVA ROC ORIGINAL
cortes <- 0:5
POS <- matrix(0, nrow = length(cortes), ncol = 2)
for (i in 1:length(cortes)) {
   obj <- vp_fp(status, teste, cortes[i])
   POS[i, ] <- c(obj$fp, obj$vp)
}
POS <- POS[!duplicated(POS), ]
ind <- order(POS[,1], POS[,2])
POS <- POS[ind, ]

# CÁLCULO DA DISTÂNCIA HORIZONTAL E DA VERTICAL
m <- length(status) - sum(status)  # NÚMERO DE INDIVÍDUOS NÃO DOENTES
n <- sum(status)                   # NÚMERO DE INDIVÍDUOS DOENTES

# DISTRIBUIÇÃO EMPÍRICA (NÃO PARAMÉTRICA)
delta <- 0.05 # NÍVEL DE SIGNIFICÂNCIA
g <- 1.36/sqrt(m) # QUANTIL 0.95 DA ESTATÍSTICA KS PARA INDIVÍDUOS NÃO DOENTES
h <- 1.36/sqrt(n) # QUANTIL 0.95 DA ESTATÍSTICA KS PARA INDIVÍDUOS DOENTES

# DETERMINAÇÃO DAS FRONTEIRAS SUPERIOR E INFERIOR DA BANDA DE CONFIANÇA EMPÍRICA
front.sup <- matrix(0, nrow = nrow(POS), ncol = 2)
front.inf <- matrix(0, nrow = nrow(POS), ncol = 2)
```

```
for (i in 1:nrow(POS)) {
    front.sup[i, 1] <- POS[i, 1] - g
    front.sup[i, 2] <- POS[i, 2] + h
    front.inf[i, 1] <- POS[i, 1] + g
    front.inf[i, 2] <- POS[i, 2] - h
}

# CALCULA A PROBABILIDADE DE COBERTURA VIA AMOSTRAS BOOTSTRAP
prob1 <- cont1 <- 0
for (j in 1:M) {
    # CALCULA A MATRIZ DOS PONTOS DA CURVA ROC PARA REAMOSTRA j
    obj <- dados_boot[[j]]
    status_n <- obj$status
    teste_n <-  obj$teste
    Matriz <- matrix(0, nrow = 6, ncol = 2)
    for (i in 1:6) {
        obj <- vp_fp(status_n, teste_n, pto = cortes[i])
        Matriz[i, ] <- c(obj$fp, obj$vp)
    }

    # REMOVE AS LINHAS REPETIDAS (SE HOUVER) E ORDENA AS LINHAS RESTANTES
    Matriz <- Matriz[!duplicated(Matriz), ]
    ind <- order(Matriz[,1], Matriz[,2])
    Matriz <- Matriz[ind, ]

    # VERIFICA SE TODOS OS PONTOS DA REAMOSTRA j ESTÃO CONTIDOS NA BANDA DE CONFIANÇA EMPÍRICA
    indicador1 <- 1
    for (i in 1:(nrow(Matriz))) {
        y1_inf <- interpola(front.inf, Matriz[i, 1])
        y1_sup <- interpola(front.sup, Matriz[i, 1])
        if ((Matriz[i, 2] > y1_sup) || (Matriz[i, 2] < y1_inf)) {indicador1 <- 0; break}
    }

    # ACUMULA O NÚMERO DE CURVAS ROC QUE ESTÃO DENTRO DA BANDA DE CONFIANÇA EMPÍRICA
    cont1 <- cont1 + indicador1
}

# VALOR DA PROBABILIDADE DE COBERTURA PARA A DISTRIBUIÇÃO USADA PELO MÉTODO SJR
prob1 <- cont1/M

# GRÁFICO DA CURVA ROC ORIGINAL E DAS BANDAS DE CONFIANÇA EMPÍRICAS SJR
plot(c(-0.2, 1.2), c(-0.2, 1.2), main = "(a)", type = "n", xlab = "FP", ylab = "VP",  xlim = c(-0.2,1.2),
ylim = c(-0.2, 1.2), bty = "o")
for (i in 1:length(dados_boot)) {
    graf_roc(dados_boot[[i]]$status, dados_boot[[i]]$teste, original = FALSE)
}
graf_roc(status, teste, original = TRUE)
lines(front.inf[,1], front.inf[,2], lwd = 2, type = "l", lty = 6)
lines(front.sup[,1], front.sup[,2], lwd = 2, type = "l", lty = 6)
```

C.1.6 Método das larguras fixas (FWB)

```
# LEITURA DOS DADOS
dados  <- read.table("dados.txt", header=T)
status <- dados$d
teste  <- dados$t

# FUNÇÃO QUE CALCULA A TAXA DE FP E VP PARA UM DADO PTO DE CORTE
vp_fp <- function(status, teste, pto) {
    clasf <- factor(x = as.numeric(teste > pto), levels = c(0, 1))
    tab <- table(status, clasf)
    vp <- tab["1","1"]/sum(tab["1",])
    fp <- tab["0","1"]/sum(tab["0",])
    return(list(vp = vp, fp = fp))
}

# FUNÇÃO QUE RETORNA O GRÁFICO DA CURVA ROC PARA UM DADO VALOR DE FP
graf_roc <- function(status, teste, cortes = 0:5, original = TRUE){
   Mat <- matrix(0, nrow = length(cortes), ncol = 2)
   for (i in 1:length(cortes)) {
       obj <- vp_fp(status, teste, cortes[i])
       Mat[i, ] <- c(obj$fp, obj$vp)
   }

   # REMOVE AS LINHAS REPETIDAS (SE HOUVER) E ORDENA AS LINHAS RESTANTES
   Mat <- Mat[!duplicated(Mat), ]
   ind <- order(Mat[,1], Mat[,2])
   Mat <- Mat[ind, ]

   # CONSTRUÇÃO DA CURVA ROC
   if (original) cor = "black" else cor = gray(0.6)
   lines(Mat[,1], Mat[,2], col = gray(0.6), lwd = 1)
}

# FUNÇÃO QUE RETORNA O VALOR DE VP PARA UM DADO VALOR DE FP USANDO INTERPOLAÇÃO LINEAR
roc <- function(Mat, fp){
   Mat <- Mat[!duplicated(Mat), ]
   ind <- order(Mat[,1], Mat[,2])
   Mat <- Mat[ind, ]

   if (fp == 1)
      vp <- 1 else {
        ind_x <- max(which(Mat[ ,1] <= fp))
        x1 <- Mat[ind_x ,1]; x2 <- Mat[ind_x + 1, 1]
        y1 <- Mat[ind_x ,2]; y2 <- Mat[ind_x + 1, 2]
        vp <- y1 + (fp - x1)*(y2 - y1)/(x2 - x1)
   }
   return(vp)
}
```

```
interpola <- function(Mat, x) {
   Mat <- Mat[!duplicated(Mat), ]
   ind <- order(Mat[,1], Mat[,2])
   Mat <- Mat[ind, ]

   # CALCULA O VALOR DE y VIA INTERPOLAÇÃO LINEAR DOS PONTOS EM Mat
   if (any(which(Mat[ ,1] <= x))) {
      ind_x <- max(which(Mat[ ,1] <= x))
      if (x == Mat[ind_x, 1]) {
         y <- Mat[ind_x, 2]
      } else {
         if (ind_x == nrow(Mat)) {
            x1 <- Mat[ind_x - 1, 1]; x2 <- Mat[ind_x, 1]
            y1 <- Mat[ind_x - 1, 2]; y2 <- Mat[ind_x, 2]
            if (x2 == x1) {
                y <- y2
            } else {
                y  <- y1 + (x - x1)*(y2 - y1)/(x2 - x1)
            }
         } else {
            x1 <- Mat[ind_x, 1]; x2 <- Mat[ind_x + 1, 1]
            y1 <- Mat[ind_x, 2]; y2 <- Mat[ind_x + 1, 2]
            y  <- y1 + (x - x1)*(y2 - y1)/(x2 - x1)
         }
      }
   } else {
      x1 <- Mat[1, 1]; x2 <- Mat[2, 1]
      y1 <- Mat[1, 2]; y2 <- Mat[2, 2]
      if (x2 == x1) {
          y <- y1
      } else {
          y  <- y1 + (x - x1)*(y2 - y1)/(x2 - x1)
      }
   }
   return(y)
}

M = 10000 # NÚMERO DE AMOSTRAS BOOTSTRAP

dados_boot <- vector("list", length = M) # ARRAY DOS DADOS BOOTSTRAP
set.seed(270183)
for (j in 1:M) {
    amostra <- sample(nrow(dados), replace=T)
    dados_n <- dados[amostra,]
    status_n <- dados_n[, 1]
    teste_n  <- dados_n[, 2]
    dados_boot[[j]] <- list(status = status_n, teste = teste_n)
}
```

```
# MATRIZ DOS TP's e VP's AMOSTRADOS VIA BOOTSTRAP
VP <- matrix(0, nrow = 6, ncol = M)
FP <- matrix(0, nrow = 6, ncol = M)
cortes <- 0:5
for (i in 1:6) {
    for (j in 1:M) {
        obj_boot <- dados_boot[[j]]
        status_n <- obj_boot$status
        teste_n  <- obj_boot$teste

        # VALOR DE VP PARA A AMOSTRA BOOTSTRAP j E PTO DE CORTE i
        obj <- vp_fp(status_n, teste_n, pto = cortes[i])
        VP[i, j] <- obj$vp
        FP[i, j] <- obj$fp
    }
}

# CRIAÇÃO DA MATRIZ DE COORDENADAS DA CURVA ROC ORIGINAL
cortes <- 0:5
POS <- matrix(0, nrow = length(cortes), ncol = 2)
for (i in 1:length(cortes)) {
   obj <- vp_fp(status, teste, cortes[i])
   POS[i, ] <- c(obj$fp, obj$vp)
}
POS <- POS[!duplicated(POS), ]
ind <- order(POS[,1], POS[,2])
POS <- POS[ind, ]

m <- length(status) - sum(status)  # NÚMERO DE INDIVÍDUOS NÃO DOENTES
n <- sum(status)                   # NÚMERO DE INDIVÍDUOS DOENTES

# INCLINAÇÃO ADOTADA
b <- -sqrt(m/n)

# DISTRIBUIÇÃO EMPÍRICA (NÃO PARAMÉTRICA)
delta <- 0.05 # NÍVEL DE SIGNIFICÂNCIA

# DETERMINAÇÃO DO VALOR DE d APROPRIADO
d <- 0          # VALOR INICIAL DE d
salto <- 0.001  # SALTO
achou <- FALSE
while (!achou) {
      # INCREMENTAÇÃO DO VALOR DE d
      d <- d + salto

      # DETERMINA AS FRONTEIRAS SUPERIOR E INFERIOR DA BANDA EMPÍRICA PARA A DISTÂNCIA d FIXADA
      front.sup <- matrix(0, nrow = 6, ncol = 2) # COORDENADAS DA FRONTEIRA SUPERIOR
      front.inf <- matrix(0, nrow = 6, ncol = 2) # COORDENADAS DA FRONTEIRA INFERIOR
```

```
      for (i in 1:nrow(POS)) {
          # CURVA SUPERIOR
          front.sup[i, 1] <- POS[i, 1] - sqrt(d^2/(b^2 + 1))   # COORDENADA X
          front.sup[i, 2] <- POS[i, 2] - b*sqrt(d^2/(b^2 + 1)) # COORDENADA Y
          # CURVA INFERIOR
          front.inf[i, 1] <- POS[i, 1] + sqrt(d^2/(b^2 + 1))   # COORDENADA X
          front.inf[i, 2] <- POS[i, 2] + b*sqrt(d^2/(b^2 + 1)) # COORDENADA Y
      }

      # VERIFICA A PROBABILIDADE DE COBERTURA DA BANDA EMPÍRICA
      prob1 <- cont1 <- 0
      for (j in 1:M) {
          # CALCULA A MATRIZ DOS PONTOS DA CURVA ROC PARA REAMOSTRA j
          obj <- dados_boot[[j]]
          status_n <- obj$status
          teste_n <-  obj$teste
          Matriz <- matrix(0, nrow = 6, ncol = 2)
          for (i in 1:6) {
              obj <- vp_fp(status_n, teste_n, pto = cortes[i])
              Matriz[i, ] <- c(obj$fp, obj$vp)
          }

          # REMOVE AS LINHAS REPETIDAS (SE HOUVER) E ORDENA AS LINHAS RESTANTES
          Matriz <- Matriz[!duplicated(Matriz), ]
          ind <- order(Matriz[,1], Matriz[,2])
          Matriz <- Matriz[ind, ]

          # VERIFICA SE TODOS OS PONTOS DA REAMOSTRA j ESTÃO CONTIDOS NA BANDA DE CONFIANÇA EMPÍRICA
          indicador1 <- 1
          for (i in 1:(nrow(Matriz))) {
              y1_inf <- interpola(front.inf, Matriz[i, 1])
              y1_sup <- interpola(front.sup, Matriz[i, 1])
              if ((Matriz[i, 2] > y1_sup) || (Matriz[i, 2] < y1_inf)) {indicador1 <- 0; break}
          }

          # ACUMULA O NÚMERO DE CURVAS ROC QUE ESTÃO NA BANDA DE CONFIANÇA EMPÍRICA
          cont1 <- cont1 + indicador1
      }

      # VALOR DA PROBABILIDADE DE COBERTURA PARA A DISTRIBUIÇÃO USADA PELO MÉTODO FWB
      prob1 <- cont1/M
      print(paste("d = ", d, "pc = ", prob1))

      # VERIFICA SE O VALOR DE d COBRE COM PROBABILIDADE 1 - delta
      if (prob1 >= (1 - delta)) achou <- TRUE
}

# GRÁFICO DA CURVA ROC ORIGINAL E DAS BANDAS DE CONFIANÇA EMPÍRICAS FWB
plot(c(-0.2, 1.2), c(-0.2, 1.2), main = "(a)", type = "n", xlab = "FP", ylab = "VP",  xlim = c(-0.2,1.2),
```

```
ylim = c(-0.2, 1.2), bty = "o")
for (i in 1:length(dados_boot)) {
    graf_roc(dados_boot[[i]]$status, dados_boot[[i]]$teste, original = FALSE)
}
graf_roc(status, teste, original = TRUE)
lines(front.inf[,1], front.inf[,2], lwd = 2, type = "l", lty = 6)
lines(front.sup[,1], front.sup[,2], lwd = 2, type = "l", lty = 6)
```

Livros Grátis

(http://www.livrosgratis.com.br)

Milhares de Livros para Download:

Baixar livros de Literatura
Baixar livros de Literatura de Cordel
Baixar livros de Literatura Infantil
Baixar livros de Matemática
Baixar livros de Medicina
Baixar livros de Medicina Veterinária
Baixar livros de Meio Ambiente
Baixar livros de Meteorologia
Baixar Monografias e TCC
Baixar livros Multidisciplinar
Baixar livros de Música
Baixar livros de Psicologia
Baixar livros de Química
Baixar livros de Saúde Coletiva
Baixar livros de Serviço Social
Baixar livros de Sociologia
Baixar livros de Teologia
Baixar livros de Trabalho
Baixar livros de Turismo