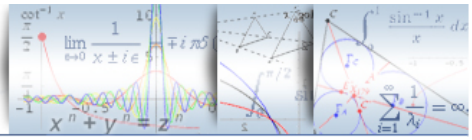
Estimação Paramétrica e Não-Paramétrica da Curva ROC

Fábio Manuel Rodrigues de Oliveira

DEPARTAMENTO DE MATEMÁTICA
FACULDADE DE CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE DE COIMBRA



Estimação Paramétrica e Não-Paramétrica da Curva ROC

Fábio Manuel Rodrigues de Oliveira

Dissertação para a obtenção do Grau de **Mestre em Matemática**

Área de Especialização em **Estatística, Optimização e Matemática Financeira**

Júri

Presidente:	Doutora Maria Esmeralda Elvas Gonçalves
Orientador:	Doutor Carlos Manuel Rebelo Tenreiro da Cruz
Vogal:	Doutora Ana Cristina Martins Rosa

Data: Agosto de 2012

Resumo

A curva ROC é um instrumento muito útil para a avaliação e comparação de diagnósticos médicos. Nesta dissertação, definimos formalmente o conceito de curva ROC, deduzindo as suas propriedades e abordamos a questão da estimação da curva ROC utilizando métodos paramétricos e não-paramétricos. Relativamente aos métodos paramétricos, começamos por estudar o estimador padrão da curva ROC, baseado no modelo binormal. Com o intuíto de contornar as suas limitações, estudamos a transformação de Box-Cox, ajustando o modelo binormal aos dados transformados. No que diz respeito à estimação não-paramétrica da curva ROC, consideramos o estimador empírico e o estimador do núcleo, estabelecendo as suas convergências locais e uniformes quase certas. A escolha das janelas para o estimador do núcleo é também abordada. Finalmente, concluímos a dissertação apresentando um estudo de simulação onde o desempenho global dos estimadores da curva ROC considerados é comparado utilizando dados simulados em seis cenários distintos.

Palavras Chave: Curva ROC, modelo binormal, estimador empírico, estimador do núcleo.

Abstract

The ROC curve is a very useful instrument for the evaluation and comparison of medical diagnostics. In this dissertation, we define formally the concept of ROC curve, deducing its properties and we address the estimation of the ROC curve by using parametric and nonparametric methods. Regarding the parametric methods, we start by studying the standard ROC curve estimator based on the binormal model. In order to circumvent its limitations, we study the Box-Cox transformation, fitting the transformed data to the binormal model. Concerning the nonparametric estimation of the ROC curve we consider the empirical and the kernel estimators and we establish their local and uniform almost sure consistency. The selection of the bandwidths for the kernel estimator is also addressed. We conclude the dissertation by presenting a simulation study where the global performance of the considered ROC curve estimators is compared by using simulated data from six different scenarios.

Keywords: ROC curve, binormal model, empirical estimator, kernel estimator.

Agradecimentos

Ao Professor Doutor Carlos Tenreiro, por ter sugerido e motivado o estudo da matéria apresentada nesta dissertação, como também pela dedicação e disponibilidade que proporcionou na minha orientação, sem as quais a concretização desta dissertação não teria sido possível.

Conteúdo

Capítulo 1

Introdução

A capacidade de poder diagnosticar uma determinada doença num indivíduo é uma ferramenta extremamente importante, não só para poder tratá-lo apropriadademente, como também para poder controlar a propagação da doença pela população, no caso de esta ser uma doença contagiosa. Um diagnóstico pode ser algo tão simples como diagnosticar uma constipação ao observar-se que o indivíduo tem o nariz congestionado, ou algo um pouco mais complicado, como diagnosticar uma infecção ao observar-se a contagem de glóbulos brancos numa recolha de sangue ou após uma biopsia. Ressonâncias magnéticas ou raios-X são também conhecidas ferramentas utilizadas na detecção de doenças, pois permitem revelar sinais físicos anómalas no paciente, que seriam invisíveis a olho nu.

Por outro lado, a prevenção também é um aspecto importante, tendo em conta o facto de que para a grande maioria das doenças, estas podem ser facilmente controladas e até curadas, se a sua detecção for feita nos estágios iniciais. Para esse efeito, tais testes têm de ser realizados regularmente o que pode ser um inconveniente, tanto a nível pessoal como económico. Um tal teste para diagnóstico será, idealmente, não-invasivo para o indivíduo, rápido e com baixos custos. Tendo em conta o facto de que o teste é utilizado na prevenção da doença, este será, na prática, aplicado a uma população maioritariamente saudável e para além disso, um diagnóstico positivo implicará, usualmente, a realização de mais testes para confirmar o primeiro diagnóstico. Assim, é tolerável um teste que não tenha uma exactidão perfeita.

Outra situação que pode também ocorrer, é a descoberta de um novo método para diagnóstico de uma determinada doença, diferente dos métodos conhecidos anteriormente. A primeira coisa que se pretenderá fazer, certamente, é avaliar a qualidade deste novo diagnóstico. A questão que se levanta portanto é a seguinte: de que forma podemos avaliar a qualidade global de um determinado diagnóstico?

Em primeiro lugar, é necessário definir de que forma se procede ao diagnóstico de uma determinada doença. Neste caso, o que o diagnóstico faz é classificar os indivíduos testados em duas classes distintas: saudável ou doente. Para efeitos de avaliação, os indivíduos vão ser provenientes de duas populações, uma população de controlo, onde todos os indivíduos são saudáveis e uma segunda população de indivíduos, dos quais já se sabe previamente estarem infectados com a doença. De seguida, procede-se a uma recolha de dados sobre os pacientes, dados estes que irão ter diferentes tipos de medida; *nominal*, em que os dados provêm de categorias (p.e. cor dos olhos), *ordinal*, em que existe uma ordem nas categorias (p.e. intensidade da dor); *discreta*, uma medição numérica que toma uma quantidade finita de valores (p.e. horas de sono diárias); medições *contínuas*, que podem tomar qualquer valor real, num conjunto limitado ou infinito (p.e. peso de um paciente). Neste último caso, o instrumento de medição terá limites de precisão, mas a medição é aceite como proveniente de uma escala contínua.

No final da recolha, o que se obtém é um vector de medições que são necessárias para o diagnóstico da doença. Essas medições vão ser todas reduzidas, através de uma transformação apropriada, a uma única variável X. Idealmente, esta transformação irá devolver valores substancialmente diferentes para indivíduos de classes diferentes, permitindo uma clara distinção entre pacientes doentes e saudáveis. Por convenção, valores mais elevados de X são indicativos da presença da doença no paciente.

De seguida, fixa-se um valor de corte, c, que vai fazer a classificação dos pacientes. Um indivíduo é classificado como doente caso o valor da variável X observado nesse indivíduo seja igual ou superior ao valor de corte. Caso contrário, considera-se saudável.

Representando por F_0 a distribuição de X em indivíduos saudáveis e por F_1 a distribuição de X em indivíduos doentes, na figura seguinte encontra-se representada uma situação típica em que são cometidos erros de diagnóstico.

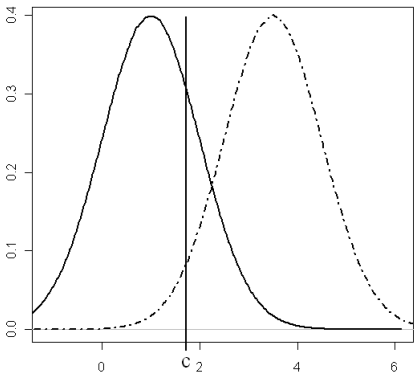


Figura 1.1: Linha - X numa população saudável. Tracejado - X numa população doente. c - Valor de corte.

Como se pode observar, as distribuições F_0 e F_1 sobrepoem-se, pelo que serão cometido erros no diagnóstico, independentemente da escolha do valor de corte. De facto, é extremamente comum para diagnósticos desta natureza eles não serem perfeitos, sendo cometido erros. Fixando o valor de c e procedendo à realização do diagnóstico, os resultados que podemos obter são:

- Positivo num paciente doente, isto é, um verdadeiro positivo (VP).
- Negativo num paciente doente, isto é, um falso negativo (FN).
- Positivo num paciente saudável, isto é, um falso positivo (FP).
- Negativo num paciente saudável, isto é, um verdadeiro negativo (VN).

De forma a podermos caracterizar o diagnóstico consoante a escolha do valor de corte fixado, temos de introduzir duas noções: a sensibilidade e a especificidade do diagnóstico. A sensibilidade define-se como sendo a probabilidade do diagnóstico devolver um resultado positivo quando é aplicado a um paciente doente. A especificidade define-se como sendo a probabilidade do diagnóstico devolver um resultado negativo quando é aplicado a um paciente saudável:

- Sensibilidade: $P(X > c|$ Paciente é doente$) = 1 - F_1(c)$.
- Especificidade: $P(X < c|$ Paciente é saudável$) = F_0(c)$.

As duas medidas anteriores estão relacionadas com a escolha do valor de corte. Um valor de corte muito alto, por exemplo, conduz a um diagnóstico muito específico, mas pouco sensível e um valor de corte baixo conduz a um diagnóstico muito sensível, mas pouco específico.

A sensibilidade pode ser estimada calculando a fracção de verdadeiros resultados positivos (FVP) devolvidos pelo diagnóstico:

$$\text{FVP} = \frac{\text{número de verdadeiros positivos}}{\text{número de pacientes doentes}} = \frac{\text{VP}}{\text{VP} + \text{FN}}.$$

A especificidade pode ser estimada calculando a fracção de verdadeiros resultados negativos (FVN) devolvidos pelo diagnóstico:

$$\text{FVN} = \frac{\text{número de verdadeiros negativos}}{\text{número de pacientes saudáveis}} = \frac{\text{VN}}{\text{VN} + \text{FP}}.$$

Na prática, o diagnóstico não é mais do que uma dicotomia, feita utilizando o valor de X e comparando-o com o valor de corte. Esse valor de corte terá de ser escolhido tendo em conta as necessidades de cada situação. As circunstâncias nas quais o diagnóstico é feito podem variar dependendo da disponibilidade de recursos ou da gravidade da doença, em que um diagnóstico mais conservador, baixando o valor de corte por exemplo, será mais razoável.

A curva ROC (Receiver Operating Characteristic) é uma ferramenta gráfica que permite descrever as características do diagnóstico para todo o conjunto de possíveis valores de corte, isto é, representa todos os possíveis valores de sensibilidade/especificidade do diagnóstico. A curva define-se parametricamente com as ordenadas a tomarem os valores da sensibilidade e as abcissas a tomarem o valor de 1−especificidade.

$$\text{ROC}(c) = \{(1 - F_0(c), 1 - F_1(c)), c \in \mathbb{R}\} \tag{1.1}$$

Utilizando os dados da Figura 1.1, podemos traçar a curva ROC para aquele caso hipotético.

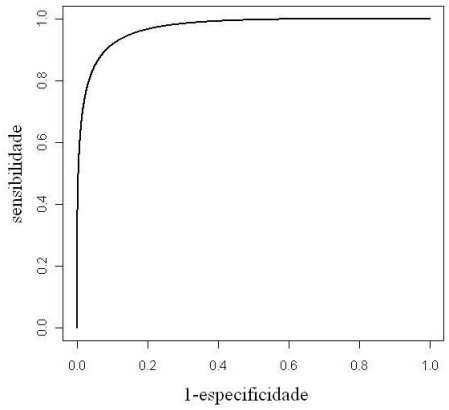


Figura 1.2: Exemplo de uma curva ROC

Um diagnóstico inútil acontece quando o parâmetro X não está relacionado de algum modo com a doença, ou seja, se tem a mesma distribuição na população saudável e na população doente. Nesse caso, a sensibilidade e a especificidade tomam sempre o mesmo valor, independentemente do valor de corte escolhido, o que equivale portanto a termos $\mathrm{ROC}(p) = p$. Por outro lado, um teste perfeito consegue separar perfeitamente os indivíduos saudáveis dos indivíduos doentes e então, existirá um valor de corte para o qual a sensibilidade e a especificidade tomam simultaneamente o valor máximo. A maioria dos diagnósticos irá produzir uma curva que se encontra entre esses dois casos especiais. Visualmente, diagnósticos com melhor qualidade têm uma curva que se aproxima do canto superior esquerdo do gráfico.

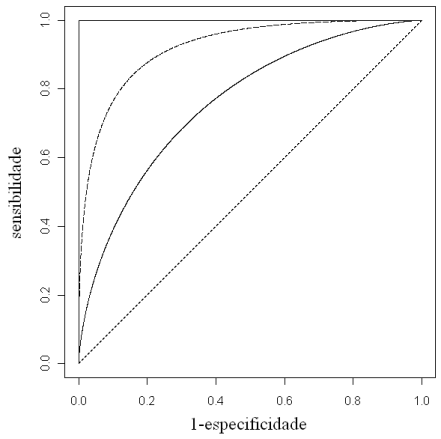


Figura 1.3: Representação de curvas ROC de diferentes diagnósticos

Neste caso, o diagnóstico representado pela curva ROC a tracejado é um diagnóstico claramente superior ao diagnóstico representado pela curva ROC representada pela curva preta, pois nenhuma escolha de valor de corte irá favorecer o diagnóstico preto com maior sensibilidade ou especificidade.

Nos próximos capítulos iremos aprofundar o estudo da curva ROC. Começaremos por reparametrizar a expressão obtida em (1.1) de forma a obtermos uma definição formal da curva ROC, dada por:

$$\mathrm{ROC}(p) = 1 - F_1\left(F_0^{-1}(1-p)\right), \quad \text{para } 0 < p < 1.$$

A partir desta definição, o nosso objectivo será o de deduzir propriedades da curva ROC. Para isso, necessitaremos de estabelecer algumas propriedades da função de distribuição F e da sua inversa generalizada, F^{-1}. Tal será feito no capítulo 2.

No capítulo 3 estudamos o modelo central da teoria da curva ROC: o modelo binormal. Daremos assim os primeiros passos na estimação paramétrica de $\mathrm{ROC}(p)$. Também iremos apresentar a transformação de Box-Cox, uma transformação que nos permitirá aplicar a estimação paramétrica a dados que, à partida, não se enquadrariam no modelo binormal.

No capítulo 4 iremos abordar dois métodos não-paramétricos para a estimação da curva ROC. Iniciaremos com o estudo do estimador empírico, $\hat{\mathrm{ROC}}(p)$, da curva ROC. Definido o estimador, apresentaremos resultados de convergência, provando que $\hat{\mathrm{ROC}}(p)$ converge pontualmente e uniformemente para $\mathrm{ROC}(p)$. Após isso, prosseguiremos com o estudo do estimador do núcleo, $\tilde{\mathrm{ROC}}(p)$ da curva ROC, apresentando resultados de convergência pontual e uniforme para este estimador. Para além disso, será analisada a escolha das janelas óptima necessárias na construção do estimador do núcleo.

Finalmente, apresentamos no capítulo 5 um estudo de simulação, onde iremos comparar a eficácia de cada um dos estimadores estudados.

No decorrer desta dissertação, denotaremos por $X_n \xrightarrow{q.c.} X$ uma sucessão de variáveis aleatórias X_n convergentes para X com probabilidade 1, quando $n \to \infty$. Denotaremos também por $o(1)$ uma sucessão de variáveis aleatórias que converge para 0 e por $O(1)$ se for limitada em probabilidade.

Capítulo 2

A Curva ROC e algumas das suas propriedades

2.1. Definição de curva ROC

Neste capítulo, iremos definir formalmente a curva ROC e pegando nessa definição, iremos também estabelecer as propriedades da curva ROC. A partir deste ponto, vamos denotar por X_0 e X_1 as observações da variável X provenientes de populações saudáveis e doentes, respectivamente, onde X_0 tem distribuição F_0 e, por seu lado, X_1 terá distribuição F_1. No capítulo anterior, introduzimos a curva ROC através de uma representação paramétrica. Se pegarmos nessa representação e lhe aplicarmos uma reparametrização, obtemos uma expressão que nos permite formalizar uma definição para a curva ROC:

Definição 2.1. *A curva ROC associada ao diagnóstico de uma determinada doença é dada pela expressão*

$$\mathrm{ROC}(p) = 1 - F_1\left(F_0^{-1}(1-p)\right), 0 < p < 1.$$

onde F_0^{-1} é a inversa generalizada da função de distribuição F_0, também conhecida como função quantil, definida por

$$F_0^{-1}(y) = \inf\{x \in \mathbb{R} : F_0(x) \geq y\}, \;\; 0 \leq y \leq 1.$$

Repare-se que a definição da curva ROC utiliza a função quantil de F_0 e por isso é necessário conhecer algumas das propriedades de F_0^{-1} para podermos deduzir propriedades da curva ROC e, mais à frente, propriedades do seu estimador.

2.2. Algumas propriedades de F e F^{-1}

As propriedades que apresentamos a seguir são demonstradas em Shorack e Wellner [10] (p. 5-8).

Proposição 2.2. *Seja F uma função de distribuição. A função $F^{-1}(t)$, $0 < t < 1$, é não-decrescente, contínua à esquerda e satisfaz*

$$F(x) \geq t \text{ se e só se } x \geq F^{-1}(t).$$

Proposição 2.3. *Para toda a função de distribuição F e a sua inversa generalizada F^{-1}, verifica-se que:*

a) $F^{-1}(F(x)) \leq x, \ \forall x: \ -\infty < x < +\infty$

b) $P\left(F^{-1}(F(X)) \neq X\right) = 0$

onde X é uma v.a.r. com distribuição F.

Propriedade 2.4. *Se F é uma função de distribuição, então*

a) F é contínua se e só se F^{-1} for estritamente crescente,

b) F é estritamente crescente se e só se F^{-1} for contínua.

Propriedade 2.5. *Seja Y uma variável aleatória com função de distribuição F contínua. Então, $U = F(Y)$ possui uma distribuição uniforme sobre o intervalo $[0,1]$, isto é, $U \sim \mathbb{U}[0,1]$.*

Demonstração. Pela Proposição 2.3, para $x \in \mathbb{R}$, temos

$$\begin{aligned} P(U \leq t) &= P(F(Y) \leq t) \\ &= P(Y \leq F^{-1}(t)) \\ &= F(F^{-1}(t)) \\ &\leq t \end{aligned}$$

Assim, concluímos que $U \sim \mathbb{U}[0,1]$. □

Proposição 2.6. *Seja U uma variável uniforme sobre o intervalo $]0,1[$ e F uma função de distribuição fixa à partida. Então a variável $X = F^{-1}(U)$ tem função de distribuição F.*

Demonstração. Pela Proposição 2.2, para $x \in \mathbb{R}$, temos

$$\begin{aligned} P(X \leq x) &= P(F^{-1}(U) \leq x) \\ &= P(U \leq F(x)) \\ &= F(x), \end{aligned}$$

o que prova o pretendido. □

2.3. Algumas propriedades de ROC(p)

Agora que vimos algumas das propriedades da função quantil, podemos prosseguir o nosso estudo e demonstrar algumas propriedades da curva ROC.

Proposição 2.7. *A curva ROC é uma função não decrescente definida em* $]0,1[$ *e com valores no intervalo* $[0,1]$.

Demonstração. Suponhamos que temos dois quaisquer pontos, p e q, pertencentes a $]0,1[$ com $p < q$.

$$\begin{aligned}\mathrm{ROC}(q) - \mathrm{ROC}(p) &= \big(1 - F_1(F_0^{-1}(1-q)\big) - \big(1 - F_1(F_0^{-1}(1-p))\big) \\ &= F_1(F_0^{-1}(1-p)) - F_1(F_0^{-1}(1-q))\end{aligned}$$

Da Proposição 2.2 e como por hipótese $p < q$, deduz-se que $F_0^{-1}(1-p) \geq F_0^{-1}(1-q)$. Como F_1 é uma função de distribuição, e portanto não-decrescente, deduz-se assim que $\mathrm{ROC}(q) \geq \mathrm{ROC}(p)$. □

Proposição 2.8. *A curva ROC é invariante para transformações estritamente crescentes dos dados.*

Demonstração. Suponhamos então que $h : \mathbb{R} \to \mathbb{R}$ é uma transformação estritamente crescente e denotemos por G_0 a função de distribuição dos dados transformados provenientes de uma população saudável e por G_1 a função de distribuição dos dados provenientes da população doente. As expressões das curvas ROC, antes e após transformação, são dadas, respectivamente, por

$$\mathrm{ROC}_X(p) = 1 - F_1(F_0^{-1}(1-p)) \text{ e } \mathrm{ROC}_{h(X)}(p) = 1 - G_1(G_0^{-1}(1-p))$$

Comecemos por provar que, para todo $p \in]0,1[$, se $x = F_0^{-1}(1-p)$ então $h(x) = G_0^{-1}(1-p)$. Suponhamos então que, para um p fixo, $x = F_0^{-1}(1-p)$. Da definição de inversa generalizada temos que,

$$x = \inf\{u \in \mathbb{R} : F_0(u) \geq 1-p\}.$$

Daqui deduzimos que,

$$G_0(h(x)) = P(h(X_0) \leq h(x)) = P(X_0 \leq x) \geq (1-p).$$

Suponhamos que existe $z \neq x$ tal que $G_0^{-1}(1-p) = h(z)$, isto é,

$$h(z) = \inf\{u \in \mathbb{R} : G_0(u) \geq 1-p\}.$$

Repetindo a mesmo dedução,

$$(1-p) \leq G_0\left(h(z)\right) = P(h(X_0) \leq h(z)) = P(X_0 \leq z) = F_0(z).$$

Portanto, temos por um lado que $x < z$ e por outro que $h(z) < h(x)$, o que nos leva a uma contradição pois por hipótese, h é uma transformação estritamente crescente. Demonstrámos assim que se $x = F_0^{-1}(1-p)$ então $h(x) = G_0^{-1}(1-p)$, para todo $p \in]0,1[$. Pegando na conclusão que acabámos de fazer, procedemos à seguinte dedução

$$\begin{aligned} G_1\left(G_0^{-1}(1-p)\right) &= G_1\left(h(F_0^{-1}(1-p))\right) \\ &= P\left(h(X_1) \leq h(F_0^{-1}(1-p))\right) \\ &= P\left(X_1 \leq h^{-1}[h(F_0^{-1}(1-p))]\right) \\ &= P\left(X_1 \leq F_0^{-1}(1-p)\right) \\ &= F_1\left(F_0^{-1}(1-p)\right). \end{aligned}$$

Assim, fica provado que $\mathrm{ROC}_X(p) = \mathrm{ROC}_{h(X)}(p)$. □

Como foi já referido anteriormente, a curva ROC é uma ferramenta muito útil na comparação de diferentes diagnósticos. De facto, a curva ROC proporciona uma referência visual relativamente fácil de interpretar. No entanto, na prática nem sempre é fácil conseguir comparar a qualidade global de um diagnóstico relativamente a outro, olhando para as suas respectivas curvas. Para esse efeito, seria conveniente termos uma medida de avaliação que nos desse uma ideia da qualidade global do diagnóstico.

2.4. Área sob a curva ROC

O cálculo da *Área sob a Curva* (AUC) proporciona-nos uma medida que permite avaliar globalmente a qualidade de um diagnóstico e é dada pela seguinte expressão:

$$\mathrm{AUC} = \int_0^1 \mathrm{ROC}(p)\, dp$$

O facto de termos uma medida numérica permite-nos comparar mais facilmente os diagnósticos. Intuitivamente, na presença de vários diagnósticos, o melhor diagnóstico, em termos globais, será o diagnóstico com o valor da área sob a curva superior.

Um diagnóstico perfeito, com uma curva ROC perfeita, terá AUC = 1. Do lado oposto, um diagnóstico irrelevante, isto é, $\mathrm{ROC}(p) = p$, terá AUC = 0.5. Na

prática, os diagnósticos terão valores entre 0 e 1. No caso de termos dois diagnósticos diferentes, A e B, por exemplo, e se souber previamente que o diagnóstico A é uniformemente superior ao diagnóstico B, ou seja,

$$\mathrm{ROC}_A(p) \geq \mathrm{ROC}_B(p),\ \forall p \in]0,1[,$$

então

$$\mathrm{AUC}_A \geq \mathrm{AUC}_B.$$

No entanto, o recíproco não é necessariamente verdade.

A área sob a curva tem uma segunda interpretação bastante interessante. De facto, escolhendo-se aleatóriamente um indivíduo da população saudável e um indivíduo da população doente, o valor de AUC exprime-se como sendo a probabilidade do parâmetro X ser superior no indivíduo doente. Mais simplesmente:

Teorema 2.9. *A área sob a curva pode ser dada pela seguinte expressão*

$$\mathrm{AUC} = P(X_1 > X_0),$$

Demonstração. Representando por U uma variável uniforme sobre o intervalo $]0,1[$, temos

$$\begin{aligned}
\mathrm{AUC} &= \int_0^1 \mathrm{ROC}(p)\, dp \\
&= \int_0^1 \Big(1 - F_1\big(F_0^{-1}(1-p)\big)\Big)\, dp \\
&= 1 - E\Big[F_1\big(F_0^{-1}(U)\big)\Big].
\end{aligned}$$

Usando agora a Proposição 2.6, podemos escrever:

$$\begin{aligned}
\mathrm{AUC} &= 1 - E\big[F_1(X_0)\big] \\
&= 1 - E\Big[P_1\big(]-\infty, X_0]\big)\Big] \\
&= E\Big[P_1\big(]X_0, +\infty[\big)\Big] \\
&= \int P_1\big(]x, +\infty[\big) dF_0(x) \\
&= \iint_{]x,+\infty[} dF_1(y) dF_0(x) \\
&= \iint_{x<y} dF_1(y) dF_0(x) \\
&= P(X_0 < X_1).
\end{aligned}$$

□

Capítulo 3

O Modelo Binormal e Estimação da Curva ROC

3.1. Modelo Binormal

Neste capítulo, o nosso estudo vai concentrar-se no estudo do modelo binormal da curva ROC e, mais adiante, iniciaremos o estudo da estimação da curva ROC, introduzindo a estimação paramétrica. Da mesma forma que a distribuição normal é um modelo clássico na Estatística Inferêncial, a curva ROC derivada a partir de dados que seguem distribuições normais é o modelo central no estudo da curva ROC a que chamamos modelo binormal. Como veremos mais à frente, este modelo dá uma boa aproximação para uma grande variedade de distribuições de dados, para além da distribuição normal.

Suponhamos então que na população de indivíduos saudáveis a variável X segue uma distribuição normal, isto é, $X_0 \sim N(\mu_0, \sigma_0^2)$ e que na população de indivíduos doentes, a variável X segue também uma distribuição normal, digamos, $X_1 \sim N(\mu_1, \sigma_1^2)$. Denotemos por Φ a função de distribuição da lei normal centrada e reduzida e por Φ^{-1} a sua inversa generalizada.

Da Definição 2.1 a curva ROC é neste caso dada por

$$\begin{aligned}
\mathrm{ROC}(p) &= 1 - F_1\big(F_0^{-1}(1-p)\big) \\
&= \Phi\left(\frac{\mu_1 - (\mu_0 - \sigma_0\Phi^{-1}(p))}{\sigma_1}\right) \\
&= \Phi\left(\frac{\mu_1 - \mu_0 + \sigma_0\Phi^{-1}(p)}{\sigma_1}\right) \\
&= \Phi\left(\frac{\mu_1 - \mu_0}{\sigma_1} + \left(\frac{\sigma_0}{\sigma_1}\right)\Phi^{-1}(p)\right), \quad 0 < p < 1.
\end{aligned}$$

Também podemos deduzir uma nova expressão para a fórmula da área sob a curva se assumirmos a binormalidade dos dados. De acordo com o Teorema 2.9 e

tendo em conta que $X_1 - X_0 \sim N(\mu_1 - \mu_0, \sigma_1^2 + \sigma_0^2)$ temos

$$\begin{aligned} \text{AUC} = 1 - P(X_1 - X_0 \leq 0) &= 1 - \Phi\left(\frac{-\mu_1 + \mu_0}{\sqrt{\sigma_1^2 + \sigma_0^2}}\right) \\ &= \Phi\left(\frac{\mu_1 + \mu_0}{\sqrt{\sigma_1^2 + \sigma_0^2}}\right). \end{aligned}$$

Se soubermos que os dados provêm de populações normais, a estimação da curva ROC é relativamente simples, bastando estimar a média e a variância de cada uma das populações. Tendo uma amostra de tamanho n_0, de uma população de indivíduos saudáveis e uma amostra de tamanho n_1, de uma população de indivíduos doentes, temos que

$$X_{0i} \sim N(\mu_0, \sigma_0^2), i \in \{1, ..., n_0\},$$

e da mesma forma,

$$X_{1j} \sim N(\mu_1, \sigma_1^2), j \in \{1, ..., n_1\}.$$

Como estimadores das médias e variância das duas populações, vamos utilizar a média amostral e a variância amostral que são os estimadores da máxima verosimilhança da média e da variância (ver Gonçalves e Mendes-Lopes [3] (p. 161-162)). Supondo então que temos uma amostra proveniente de uma população saudável de tamanho n_0, a média amostral e o desvio padrão amostral são então dados por

$$\hat{\mu}_0 = \frac{1}{n_0}\sum_{i=1}^{n_0} X_{0i} \quad \text{e} \quad \hat{\sigma}_0 = \sqrt{\frac{1}{n_0}\sum_{i=1}^{n_0}(X_{0i} - \hat{\mu}_0)^2}.$$

Da mesma forma, supondo que temos uma amostra proveniente de uma população doente de tamanho n_1, a média amostral e o desvio padrão amostral são dados por

$$\hat{\mu}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} X_{1i} \quad \text{e} \quad \hat{\sigma}_1 = \sqrt{\frac{1}{n_1}\sum_{i=1}^{n_1}(X_{1i} - \hat{\mu}_1)^2}.$$

Podemos finalmente dar uma expressão para o estimador da curva ROC segundo o modelo binormal:

$$\hat{\text{ROC}}_N(p) = \Phi\left(\frac{\hat{\mu}_1 - \hat{\mu}_0}{\hat{\sigma}_0} + \left(\frac{\hat{\sigma}_0}{\hat{\sigma}_1}\right)\Phi^{-1}(p)\right), \ \ 0 < p < 1.$$

A valor da área sob a curva, por sua vez, é estimada por

$$\hat{\text{AUC}}_N = \Phi\left(\frac{\hat{\mu}_1 + \hat{\mu}_0}{\sqrt{\hat{\sigma}_1^2 + \hat{\sigma}_0^2}}\right).$$

3.2. Transformação de Box-Cox

Já foi estabelecido anteriormente que a curva ROC é invariante para transformações estritamente crescentes dos dados. No entanto, apesar de poderem existir transformações estritamente crescentes que conservam a normalidade dos dados, nem todas o fazem. Seguindo este raciocínio, se aplicarmos uma transformação estritamente crescente a dados normais, o resultado dessa transformação podem ser dados que não seguem distribuições normais, mas que no entanto, possuem a mesma curva ROC que os dados normais iniciais. O que daqui se conclui, é facto de dados normais partilharem a mesma curva ROC com dados que não o são. Esta propriedade torna-se particularmente interessante se fizermos o raciocínio contrário, isto é, aplicarmos uma transformação adequada a dados não-normais de modo a que resultem dados normais ou aproximadamente normais, para podermos usar o modelo binormal para estimar a curva ROC.

A transformação de Box-Cox [2] é uma transformação paramétrica na qual se ajustam as variáveis a um modelo normal. O parâmetro, λ, que define a transformação será estimado a partir dos dados. A transformação define-se da seguinte forma,

$$\psi(y) = \begin{cases} \dfrac{y^\lambda - 1}{\lambda}, & \lambda \neq 0 \\ \log(y), & \lambda = 0. \end{cases}$$

Esta transformação é monótona crescente para dados estritamente positivos. Assim, vamos ter de admitir que os dados X são estritamente positivos, mas tal não constitui uma restrição importante na aplicação da transformação visto ser um facto que habitualmente ocorre em stiuações práticas. Se ainda assim estivermos na presença de um conjunto de dados em que as variáveis não são todas estritamente positivas, podemos resolver o problema adicionando um valor constante adequado a cada uma das variáveis, de modo a que todas tenham um valor estritamente positivo.

Denotemos por $X'_{0i},\ \ i \in \{1, ..., n_0\}$ as observações provenientes de uma amostra de uma população saudável após transformação e por $X'_{1j},\ \ j \in \{1, ..., n_1\}$ as observações provenientes de uma amostra de uma população doente após transformação, isto é,

$$X'_{0i} = \psi(X_{0i}) = \frac{X_{0i}^\lambda - 1}{\lambda} \quad \text{e} \quad X'_{1j} = \psi(X_{1j}) = \frac{X_{1j}^\lambda - 1}{\lambda}.$$

Após a transformação, passaremos a assumir que as variáveis seguem uma dis-

tribuição normal, ou seja, $X'_{0i} \sim N(\mu_0, \sigma_0^2)$ e $X'_{1j} \sim N(\mu_1, \sigma_1^2)$. A estimação é então feita segundo o modelo binormal, utilizando os dados transformados.

Agora que vimos de que forma a transformação será utilizada para a estimação da curva ROC, é necessário deduzirmos um método que nos permita determinar qual é a transformação adequada a utilizar para cada amostra. De facto, tratando-se de uma transformação paramétrica, em cada caso será preciso calcular o λ que melhor ajuste os dados a uma distribuição normal. Não só isso, mas também é necessário termos em conta que os dados são provenientes de duas populações diferentes e que a transformação será aplicada a cada uma das amostras das populações. Portanto o parâmetro λ que procuramos terá de ser o que melhor ajusta os dados a distribuições normais, de forma simultânea.

O método que foi escolhido para resolver o nosso problema é o método da máxima verosimilhança, proposto por Zou e Hall [13]. Para tal, vamos necessitar de calcular uma função de verosimilhança, cujo máximo nos dará o valor do parâmetro λ a utilizar na transformação. A função de verosimilhança que utilizamos é a do conjunto de dados provenientes de duas amostras de X_0 e X_1, $(x_{01}, ..., x_{0n_0}, x_{11}, ..., x_{1n_1})$, na hipótese dos dados transformados $(x'_{01}, ..., x'_{0n_0}, x'_{11}, ..., x'_{1n_1})$ serem normais. Denotando a densidade das variáveis transformadas por $f_{X'_0}$ e por f_{X_0} a densidade das variáveis originais, temos então

$$\mathscr{L}(\lambda|x'_{01}, ..., x'_{0n_0}, x'_{11}, ..., x'_{1n_1}) = f(x'_{01}, ..., x'_{0n_0}, x'_{11}, ..., x'_{1n_1}|\lambda).$$

Como as amostras de cada uma das populações são independentes, temos que

$$\begin{aligned}\mathscr{L}(\lambda|x'_{01}, ..., x'_{0n_0}, x'_{11}, ..., x'_{1n_1}) &= f_{X'_0}(x'_{01}, ..., x'_{0n_0}|\lambda) \cdot f_{X'_1}(x'_{11}, ..., x'_{1n_1}|\lambda) \\ &= \mathscr{L}_0(\lambda|X'_0) \cdot \mathscr{L}_1(\lambda|X'_1).\end{aligned}$$

Logaritmizando obtemos,

$$\log \mathscr{L}(\lambda|X'_0, X'_1) = \log \mathscr{L}_0(\lambda|X'_0) + \log \mathscr{L}_1(\lambda|X'_1),$$

e portanto, podemos construir uma função de verosimilhança correspondente a uma amostra de cada uma das populações e somá-las de seguida. Como vimos anteriormente, X'_{0i} exprime-se através de uma função de λ. Este facto, associado à assumpção de que a variável X'_{0i} segue uma distribuição normal, permite-nos construir uma função de verosimilhança, através da qual estimaremos λ. Assim,

$$\psi(X_{0i}) = \frac{X_{0i}^{\lambda} - 1}{\lambda} = X'_{0i} \xrightarrow{g} X_{0i}$$

onde g é uma transformação. Assim,

$$\begin{aligned} f_{X_0}(x) &= (f_{X_0'} \circ g^{-1})(x) \cdot |det\ J_{g^{-1}}(x)| \\ &= f_{X_0'}(\psi(x)) \cdot |det\ J_{\psi}(x)| \end{aligned}$$

onde ψ denota a transformação de Box-Cox. Esta expressão que acabámos de deduzir proporciona uma função de verosimilhança que depende do parâmetro λ. Suponhamos então que temos uma amostra de tamanho n_0 de X_0', onde $X_0' \sim N(\mu_0', \sigma_0'^2)$.

A função de verosimilhança é dada por

$$\begin{aligned} \mathscr{L}_0(\lambda, \mu_0', \sigma_0'^2 | X_0') &= \left(\frac{1}{2\pi\sigma_0'^2}\right)^{\frac{n_0}{2}} \exp\left(-\frac{1}{2\sigma_0'^2}\sum_{i=1}^{n_0}\left(x_{0i}' - \mu_0'\right)^2\right) \\ &= \left(\frac{1}{2\pi\sigma_0'^2}\right)^{\frac{n_0}{2}} \exp\left(-\frac{1}{2\sigma_0'^2}\sum_{i=1}^{n_0}\left(\frac{x_{0i}^{\lambda}-1}{\lambda} - \mu_0'\right)^2\right)\prod_{i=1}^{n_0}\left(\frac{x_{0i}^{\lambda}-1}{\lambda}\right)' \\ &= \left(\frac{1}{2\pi\sigma_0'^2}\right)^{\frac{n_0}{2}} \exp\left(-\frac{1}{2\sigma_0'^2}\sum_{i=1}^{n_0}\left(\frac{x_{0i}^{\lambda}-1}{\lambda} - \mu_0'\right)^2\right)\prod_{i=1}^{n_0}\left(x_{0i}^{(\lambda-1)}\right). \end{aligned}$$

Esta função depende de três parâmetros: λ, μ_0' e σ_0'. No entanto, se utilizarmos estimativas para μ_0' e σ_0', é possível exprimir esses parâmetros como funções de λ. De facto, utilizando as fórmulas da média e da variância amostrais, após logaritmizarmos, obtemos então:

$$\log \mathscr{L}_0(\lambda|X_0') = \frac{n_0}{2}\log\left(\frac{1}{2\pi\sigma_0'^2}\right) - \frac{1}{2\sigma_0'^2}\sum_{i=1}^{n_0}\left(\frac{x_{0i}^{\lambda}-1}{\lambda} - \mu_0'\right)^2 + (\lambda-1)\sum_{i=1}^{n_0}\log(x_{0i}),$$

onde $\mu_0' = \frac{1}{n_0}\sum_{i=1}^{n_0}\left(\frac{x_{0i}^{\lambda}-1}{\lambda}\right)$ e $\sigma_0'^2 = \frac{1}{n_0}\sum_{i=1}^{n_0}\left(\frac{x_{0i}^{\lambda}-1}{\lambda} - \mu_0'\right)^2$. Se procedermos da forma análoga para a população doente, obtemos a função de verosimilhança:

$$\log \mathscr{L}_1(\lambda|X_1') = \frac{n_1}{2}\log\left(\frac{1}{2\pi\sigma_1'^2}\right) - \frac{1}{2\sigma_1'^2}\sum_{j=1}^{n_1}\left(\frac{x_{1j}^{\lambda}-1}{\lambda} - \mu_1'\right)^2 + (\lambda-1)\sum_{j=1}^{n_1}\log(x_{1j}).$$

Finalmente, a função de verosimilhança utilizada na estimação do parâmetro λ é dada por

$$\log \mathscr{L}(\lambda|X_0', X_1') = \log \mathscr{L}_0(\lambda|X_0') + \log \mathscr{L}_1(\lambda|X_1').$$

Agora que possuimos todas as ferramentas de que necessitamos, podemos recapitular enunciado os passos da estimação. Primeiro, estima-se o valor do parâmetro λ. De seguida, aplicamos a transformação aos dados. Finalmente, calculam-se os valores das médias e variâncias amostrais e aplicamos esses valores ao modelo binormal, obtendo assim, uma estimativa da curva ROC.

3.3. Sobre o valor de corte c

Apesar de este não ser o objectivo do nosso estudo, vale a pena fazer uma referência quanto à escolha do valor de corte. Na realização do diagnóstico, a escolha do valor de corte pode ser feita não só com base na sensibilidade e na especificidade que se pretende obter no diagnóstico, mas também utilizando duas outras medidas: o valor preditivo positivo (VPP), que determina a proporção de resultados positivos correctos e o valor preditivo negativo (VPN), que determina a proporção de resultados negativos correctos, e que na prática são escolhidos consoante a prevalência da doença que se está a diagnosticar na população. Estas medidas são definidas por:

- VPP= $P(\text{Paciente ser doente}|X \geq c)$
- VPN= $P(\text{Paciente ser saudável}|X < c)$

Peguemos na definição de VPP. Denotando por D o acontecimento 'o paciente ser doente' e procedendo a uma aplicação directa do teorema de Bayes, obtemos

$$\begin{aligned}\text{VPP} &= P\left(D|X \geq c\right) \\ &= \frac{P(X \geq c|D)P(D)}{P(X \geq c|D)P(D) + P(X \geq c|\overline{D})P(\overline{D})} \\ &= \frac{\text{(sensibilidade)(prevalência)}}{\text{(sensibilidade)(prevalência)} + \text{(1-sensibilidade)(1-prevalência)}}.\end{aligned}$$

Usando o mesmo raciocínio para VPN, obtemos

$$\text{VPP} = \frac{\text{(especificidade)(1-prevalência)}}{\text{(especificidade)(1-prevalência)} + \text{(1-especificidade)(prevalência)}}.$$

Como se pode observar, os valores preditivos dependem não só da sensibilidade e da especificidade, mas também da prevalência da doença na população. De facto, um valor preditivo positivo muito baixo pode ser resultado de o diagnóstico não qualificar convenientemente o doente, mas pode também ser o resultado de uma prevalência baixa da doença na população em causa. A sensibilidade e a especificidade são medidas mais eficazes na avaliação do diagnóstico, mas na prática, pode ser mais interessante saber o quão provável a doença está de facto presente, quando o resultado do diagnóstico é positivo.

Suponhamos que temos dois diagnósticos para a mesma doença. Em ambos os casos, a variável X segue distribuições normais na população saudável e na população doente e cujas densidades são as seguintes, em cada diagnóstico:

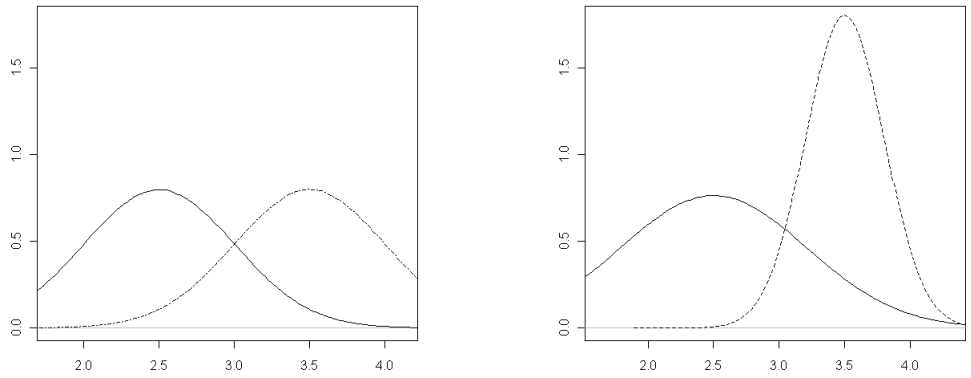


Figura 3.1: Densidades de X_0 e X_1 para o primeiro (esquerda) e segundo (direita) diagnósticos.

Sendo as distribuições em cada uma das populações conhecidas, para os dois diagnósticos, podemos traçar as suas respectivas curvas ROC e sobrepô-las, para comparação.

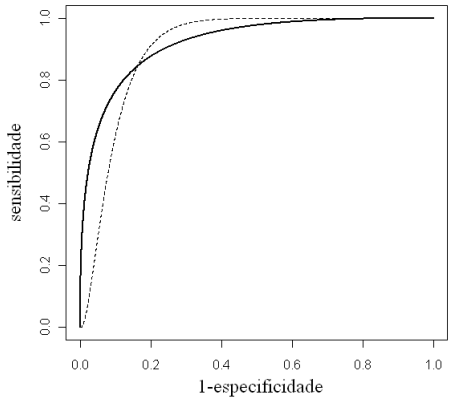


Figura 3.2: Curvas ROC do primeiro (linha contínua) e segundo (tracejado) diagnósticos.

Como podemos ver, nenhum dos diagnósticos é claramente superior ao outro. De facto, se pretendermos escolher um valor de corte de modo a ter um diagnóstico mais sensível, então o segundo diagnóstico será o mais adequado. Por outro lado, se pretendermos escolher um valor de corte de modo a ter um diagnóstico mais específico, então o primeiro será o mais adequado. Suponhamos agora que estes diagnósticos servem para detectar uma doença que tem uma prevalência de 30%. Podemos traçar as curvas ROC de cada um dos diagnósticos e respectivos valores preditivos.

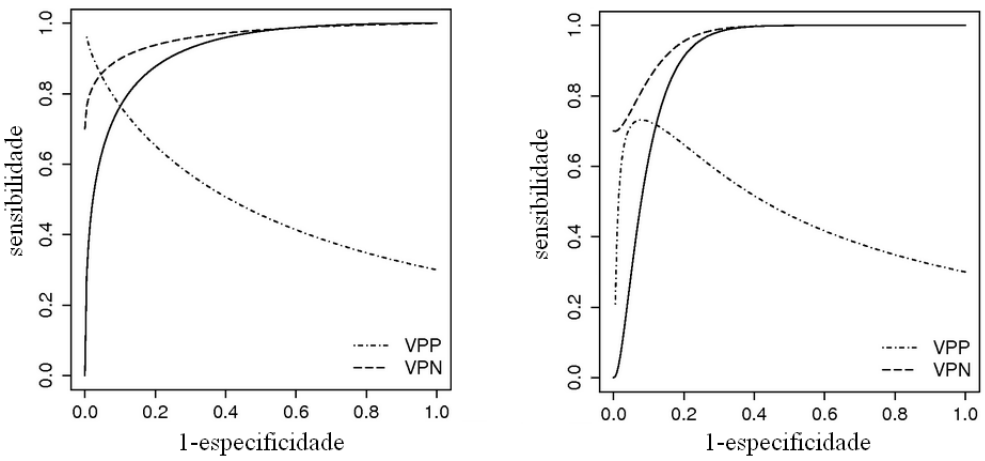


Figura 3.3: Esquerda: Diagnóstico 1. Direita: Diagnóstico 2

Observando a figura, para tal prevalência da doença, se pretendermos fazer a escolha do valor de corte, de modo a obter o maior valor preditivo positivo, então o primeiro dignóstico será claramente superior ao segundo, apesar de qualitativamente isso não ser o caso. Este exemplo serviu para mostrar que a escolha do valor de corte que define o diagnóstico na prática, não é de maneira alguma trivial e que nem sempre é feita tendo em conta a qualidade do diagnóstico.

Capítulo 4

Estimação não-paramétrica da curva ROC

Uma segunda abordagem que podemos tomar no âmbito da estimação da curva ROC é a abordagem não-paramétrica. No caso estudado no capítulo anterior, o objectivo era adequar o modelo binormal aos dados, aplicando uma transformação adequada e, se necessário, estimando os parâmetros necessários. Neste capítulo, não iremos fazer assumpções nenhumas quanto às distribuições dos dados, utilizando métodos não-paramétricos para podermos estimar a curva ROC.

4.1. O estimador empírico

O estimador empírico aplica simplesmente a definição da curva ROC aos dados da situação em estudo. Assim, para todos os valores de corte do diagnóstico, as funções de distribuição das populações são substituídas pelas funções de distribuição empíricas das amostras. Suponhamos então que temos duas amostras provenientes de populações saudáveis e doentes, com tamanhos n_0 e n_1, respectivamente. As funções de distribuição empíricas são dadas por,

$$\hat{F}_0(x) = \frac{1}{n_0}\sum_{i=1}^{n_0}\mathbb{I}(X_{0i} \leq x) \quad \text{e} \quad \hat{F}_1(x) = \frac{1}{n_1}\sum_{j=1}^{n_1}\mathbb{I}(X_{1j} \leq x), \;\; x \in \mathbb{R}$$

O estimador empírico da curva ROC é assim definido por

$$\hat{\text{ROC}}(p) = 1 - \hat{F}_1\left(\hat{F}_0^{-1}(1-p)\right), \;\; 0 < p < 1.$$

Esta função é na verdade uma função discreta pois as funções empíricas tomam apenas valores no conjunto $\{0, 1/n_0, 2/n_0, ..., 1\}$. Se não existirem resultados repetidos nos dados, a curva ROC apresentar-se-á como sendo uma função em escada com saltos horizontais de amplitude $1/n_0$, correspondentes aos resultados das observações em indivíduos saudáveis e saltos verticais de amplitude $1/n_1$, correspondentes aos resultados das observações em indivíduos doentes. Caso existam igualdades para resultados dos indivíduos saudáveis, isso vai-se traduzir em maiores saltos horizontais

e maiores saltos verticais caso se verifiquem igualdades nos resultados da população doente. Resultados iguais nos casos saudáveis e doentes, em simultâneo, resultam em segmentos de recta diagonais.

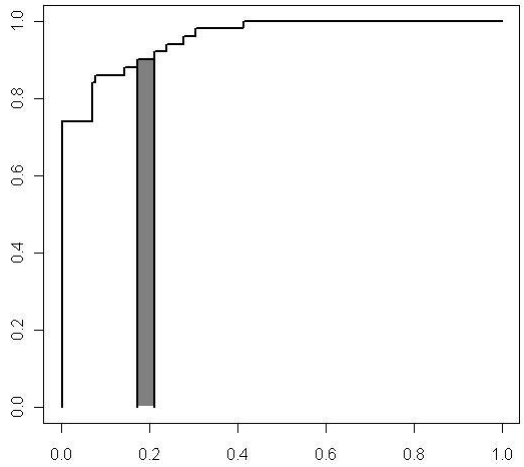


Figura 4.1: Exemplo de uma curva ROC empírica

4.1.1. Convergência pontual

Agora que definimos o estimador empírico, é necessário verificar se estamos na presença de um estimador consistente, isto é, temos de provar que $\hat{\text{ROC}}$ converge pontualmente para ROC. O primeiro resultado que vamos provar é o da convergência pontual do estimador empírico da inversa generalizada da função de distribuição.

Teorema 4.1. *Seja* $0 < p < 1$. *Se* $\xi_p = F^{-1}(p)$ *for a única solução de* $F(x-) \leq p \leq F(x)$, *onde* $F(x-) = lim_{y \uparrow x} F(y) = P(X < x)$, *então*

$$\hat{F}_n^{-1}(p) \xrightarrow{q.c.} F^{-1}(p).$$

Demonstração. Vamos começar por provar que

$$\forall \epsilon > 0, \;\; F(\xi_p - \epsilon) < p < F(\xi_p + \epsilon).$$

Peguemos na primeira desigualdade, $F(\xi_p - \epsilon) < p$. Por hipótese,

$$\xi_p = F^{-1}(p) = \inf_{x \in \mathbb{R}} \{F(x) \geq p\},$$

pela definição de função quantil. Suponhamos por absurdo que, para algum $\epsilon > 0$, temos $F(\xi_p - \epsilon) \geq p$. Visto que $\xi_p - \epsilon > \xi_p$, então

$$\xi_p - \epsilon = \inf_{x \in \mathbb{R}} \{F(x) \geq p\}.$$

No entanto, isto vai contra a nossa hipótese e portanto $F(\xi_p - \epsilon) < p$.

Demonstremos agora a segunda desigualdade. Por hipótese, $F(\xi_p-) \leq p \leq F\xi_p)$. Suponhamos por absurdo que, $p \geq F(\xi_p + \epsilon)$, para algum $\epsilon > 0$. Nesse caso,

$$F\left((\xi_p + \epsilon)^-\right) \leq F(\xi_p + \epsilon) \leq p \leq F(\xi_p) \leq F(\xi_p + \epsilon),$$

ou seja, $\xi_p + \epsilon$ é também solução de $F(x-) \leq p \leq F(x)$, o que contraria a hipótese. Para provarmos a convergência pontual da função quantil empírica, vamos ter de demonstrar que

$$\forall \epsilon > 0, \;\; P\left(|\hat{F}_m^{-1}(p) - \xi_p| < \epsilon, \forall m \geq n\right) \longrightarrow 1, n \to \infty, \tag{4.2}$$

ou equivalentemente,

$$\forall \epsilon > 0, \;\; P\left(\xi_p - \epsilon < \hat{F}_m^{-1}(p) \leq \xi_p + \epsilon, \forall m \geq n\right) \longrightarrow 1, n \to \infty.$$

Peguemos na primeira desigualdade. Aplicando a Proposição 2.2, temos:

$$\begin{aligned}
&P\left(\xi_p - \epsilon < \hat{F}_m^{-1}(p), \forall m \geq n\right) \\
&= P\left(\hat{F}_m(\xi_p - \epsilon) < p, \;\; \forall m \geq n\right) \\
&= P\left(\hat{F}_m(\xi_p - \epsilon) - F(\xi_p - \epsilon) < p - F(\xi_p - \epsilon), \;\; \forall m \geq n\right) \\
&\geq P\Big(- (p - F(\xi_p - \epsilon)) < \hat{F}_m(\xi_p - \epsilon) \\
&\quad - F(\xi_p - \epsilon) < p - F(\xi_p - \epsilon), \;\; \forall m \geq n\Big) \\
&\geq P\Big(|\hat{F}_m(\xi_p - \epsilon) - F(\xi_p - \epsilon)| < p - F(\xi_p - \epsilon), \;\; \forall m \geq n\Big).
\end{aligned}$$

Sabendo que $\hat{F} \xrightarrow{q.c.} F$ (ver Serfling [9], p. 57), temos, para algum $\delta > 0$:

$$P\left(|\hat{F}_m(\xi_p - \epsilon) - F(\xi_p - \epsilon)| < \delta, \forall m > n\right) \longrightarrow 1, \;\; n \to \infty.$$

Para além disso, já vimos que $F(\xi_p - \epsilon) < p$. Assim,

$$P\Big(|\hat{F}_m(\xi_p - \epsilon) - F(\xi_p - \epsilon)| < p - F(\xi_p - \epsilon), \;\; m \geq n\Big) \longrightarrow 1, \;\; n \to \infty,$$

e consequentemente,

$$\forall \epsilon > 0, \;\; P\left(\xi_p - \epsilon < \hat{F}_m^{-1}(p), \;\; \forall m > n\right) \longrightarrow 1, \;\; n \to \infty. \tag{4.3}$$

Agora provemos de forma semelhante que

$$\forall \epsilon > 0, \;\; P\left(\hat{F}_m^{-1}(p) < \xi_p + \epsilon, \;\; \forall m > n\right) \longrightarrow 1.$$

Uma vez mais, aplicando a Proposição 2.2 temos:

$$\begin{aligned}
&P\left(\hat{F}_m^{-1}(p) \leq \xi_p + \epsilon,\ \ \forall m > n\right) \\
&= P\left(F_m(\xi_p + \epsilon) \geq p,\ \ m \geq n\right) \\
&= P\left(F_m(\xi_p + \epsilon) - F(\xi_p + \epsilon) \geq p - F(\xi_p + \epsilon),\ \ \forall m \geq n\right) \\
&\geq P\left(|F_m(\xi_p + \epsilon) - F(\xi_p + \epsilon)| < F(\xi_p + \epsilon) - p,\ \ \forall m \geq n\right)
\end{aligned}$$

Argumentos análogos aos anteriores permitem concluir

$$P\Big(|\hat{F}_m(\xi_p + \epsilon) - F(\xi_p + \epsilon)| < F(\xi_p + \epsilon) - p,\ \ m \geq n\Big) \longrightarrow 1,\ \ n \to \infty$$

e consequentemente,

$$\forall \epsilon > 0,\ \ P\left(\hat{F}_m^{-1}(p) < \xi_p + \epsilon,\ \ \forall m > n\right) \longrightarrow 1. \tag{4.4}$$

De (4.3) e (4.4) obtemos o pretendido. □

Repare-se que, no teorema que acabámos de demonstrar, não podemos ignorar a hipótese de que $\xi_p = F^{-1}(p)$ é a única solução de $F(x-) \leq p \leq F(x)$. De facto, se tal não se verificasse, teríamos, para algum $\epsilon > 0$, $p \geq F(\xi_p + \epsilon)$, o que nos levaria a um absurdo na demonstração de (4.4).

Com este resultado, estamos em condições de estabelecer a convergência pontual de $\hat{\text{ROC}}$.

Teorema 4.5. *Seja $0 < p < 1$. Se $\xi_{1-p} = F_0^{-1}(1-p)$ for a única solução de $F_0(x-) \leq 1-p \leq F_0(x)$ e F_1 for contínua em $F_0^{-1}(1-p)$, então*

$$\hat{\text{ROC}}(p) \xrightarrow{q.c.} \text{ROC}(p).$$

Demonstração. Consideremos a seguinte majoração:

$$\begin{aligned}
\left|\hat{\text{ROC}}(p) - \text{ROC}(p)\right| &\leq \left|\hat{F}_1\left(\hat{F}_0^{-1}(1-p)\right) - F_1\left(\hat{F}_0^{-1}(1-p)\right)\right| \\
&\quad + \left|F_1\left(\hat{F}_0^{-1}(1-p)\right) - F_1\left(F_0^{-1}(1-p)\right)\right| \\
&\leq \sup_{x\in\mathbb{R}} \left|\hat{F}_1(x) - F_1(x)\right| + \left|F_1\left(\hat{F}_0^{-1}(1-p)\right) - F_1\left(F_0^{-1}(1-p)\right)\right|.
\end{aligned}$$

A primeira parcela converge para zero quase certamente, segundo o teorema de Glivenko-Cantelli (ver Serfling [9] (p.61)). Verificando-se a continuidade de F_1 em $F_0^{-1}(1-p)$, podemos aplicar o Teorema 4.1 à segunda parcela, concluindo que também converge para zero. Fica assim provada a convergência pontual do estimador. □

4.1.2. Convergência uniforme

De seguida, vamos obter a convergência uniforme do estimador empírico. Tal como no caso da convergência pontual, a convergência do estimador da curva ROC vai ser consequência da convergência do estimador empírico da função quantil, tratando-se neste caso da convergência uniforme. Comecemos então por estebelecer tal resultado.

Teorema 4.6. *Suponhamos que F é uma função de distribuição contínua. Então, para todo $\epsilon > 0$,*

$$\sup_{t\in[\epsilon,1-\epsilon]} \left|\hat{F}^{-1}(t) - F^{-1}(t)\right| \xrightarrow{q.c.} 0.$$

Demonstração. Suponhamos que temos uma amostra $Y_1, ..., Y_n$ de variáveis com distribuição F e defina-se $U_i = F(Y_i)$. Pela Proposição 2.5 sabemos que as variáveis U_i seguem a lei uniforme no intervalo $[0, 1]$. Aplicando agora a Proposição 2.3 ao nosso caso, obtemos,

$$P\left(F^{-1}(F(Y_i)) = Y_i\right) = 1 \quad \text{e portanto} \quad Y_i = F^{-1}(F(Y_i)) = F^{-1}(U_i) \quad \text{q.c.}.$$

Conclui-se que,

$$\begin{aligned}\hat{F}^{-1}(t) = Y_i, \ \text{ se } \ \frac{i-1}{n} < t \leq \frac{i}{n} \\ = F^{-1}(U_i), \ \text{ pois } F^{-1} \text{ é não decrescente} \\ = F^{-1}\left(\hat{K}^{-1}(t)\right),\end{aligned}$$

onde $\hat{K}^{-1}$ representa o estimador empírico da função quantil associada a $U_1, ..., U_n$. A dedução que foi feita leva-nos à seguinte igualdade:

$$\sup_{t\in[\epsilon,1-\epsilon]} \left|\hat{F}^{-1}(t) - F^{-1}(t)\right| = \sup_{t\in[\epsilon,1-\epsilon]} \left|F^{-1}\left(\hat{K}^{-1}(t)\right) - F^{-1}(t)\right|. \tag{4.7}$$

Analisemos a convergência de $\hat{K}^{-1}(t)$. Já sabemos que cada uma das variáveis U_i segue uma distribuição uniforme, donde se deduz que $\hat{K}(t) = t \Leftrightarrow t = \hat{K}^{-1}(t)$. Onde $\hat{K}$ representa a função de distribuição empírica associada a $U_1, ..., U_n$. Assim,

$$\sup_{t\in[\epsilon,1-\epsilon]} \left|\hat{K}^{-1}(t) - t\right| = \sup_{t\in[\epsilon,1-\epsilon]} \left|\hat{K}(t) - t\right|.$$

Pelo Teorema de Glivenko-Cantelli, temos:

$$\sup_{t\in[0,1]} \left|\hat{K}(t) - t\right| \xrightarrow{q.c.} 0,$$

e por maioria de razão, para $\epsilon > 0$,

$$\sup_{t\in[\epsilon,1-\epsilon]} \left|\hat{K}(t) - t\right| \xrightarrow{q.c.} 0.$$

Aplicando esta dedução a (4.7), concluímos que

$$\sup_{t\in[\epsilon,1-\epsilon]}\left|\hat{F}^{-1}(t)-F^{-1}(t)\right|\xrightarrow{q.c.}0.$$

□

Teorema 4.8. *Seja F_1 uma distribuição com densidade f_1 limitada. Suponhamos também que F_0 é uma distribuição contínua e estritamente crescente em $[\epsilon,1-\epsilon]$, $\epsilon>0$. Nestas condições,*

$$\sup_{p\in[\epsilon,1-\epsilon]}\left|\hat{\mathrm{ROC}}(p)-\mathrm{ROC}(p)\right|\xrightarrow{q.c.}0.$$

Demonstração. Consideremos a seguinte majoração:

$$\begin{aligned}\left|\hat{\mathrm{ROC}}(p)-\mathrm{ROC}(p)\right| &\leq \left|F_1\left(\hat{F}_0^{-1}(1-p)\right)-\hat{F}_1\left(\hat{F}_0^{-1}(1-p)\right)\right| \\ &\quad+\left|F_1\left(F_0^{-1}(1-p)\right)-F_1\left(\hat{F}_0^{-1}(1-p)\right)\right| \\ &\leq \sup_{x\in\mathbb{R}}\left|F_1(x)-\hat{F}_1(x)\right|+\left|F_1\left(F_0^{-1}(1-p)\right)-F_1\left(\hat{F}_0^{-1}(1-p)\right)\right|.\end{aligned}$$

A primeira parcela do lado direito da desigualdade converge uniformemente para zero, segundo o Teorema de Glivenko-Cantelli. Visto que F_1 tem densidade limitada, então é uniformemente contínua em $F_0^{-1}(p)$. Como F_1 tem densidade limitada, então é uma função uniformemente contínua e portanto podemos assim aplicar o Teorema 4.6 à segunda parcela, obtendo o pretendido. □

4.1.3. Área Sob a Curva

O cálculo da área sob a curva leva-nos a uma expressão interessante, quando ta

Proposição 4.9. *A área sob a curva* $\hat{\mathrm{ROC}}$ *é dada pela U-estatística de Mann-Whitney:*

$$\int_0^1\hat{\mathrm{ROC}}(p)dp=\frac{1}{n_0n_1}\sum_{j=1}^{n_0}\sum_{i=1}^{n_1}\mathbb{I}(X_{1i}>X_{0j})$$

Demonstração.

$$\begin{aligned}\hat{\mathrm{AUC}} &= \int_0^1\hat{\mathrm{ROC}}(p)dp \\ &= \int_0^1 1-\hat{F}_1\left(\hat{F}_0^{-1}(1-p)\right)dp \\ &= \int_0^1 1-\frac{1}{n_1}\sum_{i=1}^{n_1}\mathbb{I}\left(X_{1i}\leq\hat{F}_0^{-1}(1-p)\right)dp \\ &= \int_0^1\frac{1}{n_1}\sum_{i=1}^{n_1}\mathbb{I}\left(X_{1i}>\hat{F}_0^{-1}(1-p)\right)dp \\ &= \frac{1}{n_0n_1}\sum_{j=1}^{n_0}\sum_{i=1}^{n_1}\mathbb{I}(X_{1i}>X_{0j}).\end{aligned}$$

□

A demonstração deste teorema pode ser feita de uma forma intuitiva e visual utilizando a figura 4.1 como referência. Suponhamos que não existem igualdades entre observações na população doente e na população saudável. Nesse caso, cada passo horizontal de RÔC correspondente a um ponto X_{0j} adiciona uma área rectangular de valor $(1/n_0) \times \hat{F}_0(X_{0j}) = (1/n_0) \times \sum_{j=1}^{n_0} \mathbb{I}[X_{1i} > X_{0j}]/n_1$. Este resultado segue do facto da AÛC ser na verdade a soma em i dos n_0 rectângulos, como o da figura. No caso de se verificarem igualdades, entre X_{0j} e X_{1i} por exemplo, isso adiciona uma àrea triangular com a face horizontal de comprimento$1/n_0$ e comprimento $1/n_1$ na face vertical, isto é, uma área de $\frac{1}{2(n_0 n_1)}\mathbb{I}(X_{1i} = X_{0j})$.

4.2. O estimador do núcleo

Do ponto de vista prático, os dois métodos de estimação que foram estudados nos capítulos anteriores encontram-se, do ponto de vista ideológico, em pólos opostos. A abordagem paramétrica, devolve-nos uma estimativa suave da curva ROC, mas força-nos a assumir que as amostras (ou as respectivas transformações) possuem distribuições normais, o que pode levar a cometermos um erro de especificação. Por outro lado, a estimação não-paramétrica não nos obriga a fazer tais assumpções, mas a curva estimada é, na verdade, uma função em escada, que não dá realmente uma boa representação visual da curva ROC verdadeira, quando esta é contínua.

Assim, com o intuito de encontrar um método que permita estimar a curva ROC sem os inconvenientes das duas estimações referidas, Zou et al. [12] e Lloyd [6] foram os primeiros a propor uma estimativa da curva ROC usando o estimador do núcleo. A partir de duas amostras de populações saudáveis e doentes, com tamanho n_0 e n_1, respectivamente, as estimações de F_0 e F_1 são dadas por:

$$\tilde{F_0}(x) = \frac{1}{n_0}\sum_{i=1}^{n_0} L\left(\frac{x - X_{0i}}{h_0}\right) \text{ e } \tilde{F_1}(x) = \frac{1}{n_1}\sum_{j=1}^{n_1} L\left(\frac{x - X_{1j}}{h_1}\right)$$

onde $L(x) = \int_{-\infty}^{x} K(u)du$ e K é uma função densidade de probabilidade contínua e $h_0 = h_0(n_0)$, $h_1 = (n_1)$ são números reais estritamente positivos, a que chamamos janelas dos estimadores e que convergem para zero quando n_0 e n_1 tendem para infinito. Os núcleos utilizados para cada uma das estimações são os mesmos, apesar de isso não ser algo obrigatório. Ficamos então com o seguinte expressão para o estimador da curva ROC:

$$\tilde{\text{ROC}} = 1 - \tilde{F_1}\left(\tilde{F_0}^{-1}(1-p)\right).$$

Um problema subjacente a este tipo de estimador é o problema da escolha da janela. Lloyd [7] e Zhou e Harezlak [11], foram os primeiros a apresentar métodos empíricos para a escolha das janelas h_0 e h_1. No entanto, ambos os métodos trataram o problema da estimação como se estando a estimar F_0 e F_1 separadamente e não no contexto da curva ROC, o que se verificou poder gerar erros de estimação no caso das distribuições das populações serem significativamente diferentes. Hall e Hyndman [4] apresentaram uma abordagem na qual a escolha das janelas utilizadas na estimação não é feita de forma independente de cada uma das distribuições das respectivas populações.

Antes de abordarmos tal problema, o nosso primeiro passo será o de obter resultados de convergência para este estimador da curva ROC.

4.2.1. Convergência pontual

A demonstração da convergência pontual do estimador do núcleo segue os mesmo passos da demonstração para a convergência pontual do estimador empírico, mas para tal, iremos necessitar de um resultado que nos garante que o Teorema de Glivenko-Cantelli também se pode aplicar para o estimador do núcleo da função de distribuição. A demonstração que se apresentamos é da autoria de Horváth e Zhou [5].

Teorema 4.10. *Seja F uma função de distribuição contínua em $\mathbb{R}$ e $\tilde{F}(x) = \frac{1}{n}\sum_{i=1}^{n} L\left(\frac{x-X_i}{h}\right)$ o estimador de F com núcleo L e janela $h = h(n) \to 0$, $n \to \infty$. Então:*

$$\sup_{x\in\mathbb{R}} \left|\tilde{F}(x) - F(x)\right| \xrightarrow{q.c.} 0.$$

Demonstração. Visto que F é contínua em $\mathbb{R}$, é uniformemente contínua em $\mathbb{R}$. Seja $\hat{F}(x) = n^{-1}\sum_{i=1}^{n} \mathbb{I}(X_i \leq x)$. Temos,

$$\begin{aligned}
&\sup_{x\in\mathbb{R}} \left|\tilde{F}(x) - F(x)\right| \\
&\leq \sup_{x\in\mathbb{R}} \left| \int L\left(\frac{x-t}{h}\right) d\hat{F}(t) - F(x) + \int L\left(\frac{x-t}{h}\right) dF(t) - \int L\left(\frac{x-t}{h}\right) dF(t) \right| \\
&\leq \sup_{x\in\mathbb{R}} \left| \int L\left(\frac{x-t}{h}\right) d(\hat{F}(t) - F(t)) \right| + \sup_{x\in\mathbb{R}} \left| \int L\left(\frac{x-t}{h}\right) dF(t) - F(x) \right| \qquad (4.11) \\
&= C_{1,n} + C_{2,n}.
\end{aligned}$$

Comecemos por estimar C_{1_n}. Se fizermos $n \to \infty$, temos pelo Teorema de Glivenko-

Cantelli:

$$\begin{aligned} C_{1,n} &= \sup_{x\in\mathbb{R}} \left| \frac{1}{h} \int (\hat{F}(t) - F(t)) K\left(\frac{x-t}{h}\right) dt \right| \\ &= \sup_{x\in\mathbb{R}} \left| \int (\hat{F}(x-uh) - F(x-uh)) K(u) du \right| \\ &\leq \sup_{x\in\mathbb{R}} \left| \hat{F}(x) - F(x) \right| \xrightarrow{q.c.} 0. \end{aligned}$$

Analisemos agora o caso de $C_{2,n}$. Procedendo a uma mudança de variáveis, obtemos

$$\begin{aligned} C_{2,n} &= \sup_{x\in\mathbb{R}} \left| \frac{1}{h} \int F(t) K\left(\frac{x-t}{h}\right) dt - F(x) \right| \\ &= \sup_{x\in\mathbb{R}} \left| \int (F(x-uh) - F(x)) K(u) du \right| \\ &\leq \sup_{x\in\mathbb{R}} \int_{-\infty}^{-A} |F(x-uh) - F(x)| K(u) du + \sup_{x\in\mathbb{R}} \int_{A}^{\infty} |F(x-uh) - F(x)| K(u) du \\ &\quad + \sup_{x\in\mathbb{R}} \int_{-A}^{A} |F(x-uh) - F(x)| K(u) du \\ &\leq L(-A) + 1 - L(A) + \sup_{x\in\mathbb{R}} |F(x+Ah) - F(x-Ah)|. \end{aligned}$$

Se A for suficientemente grande, $L(-A)$ e $1 - L(A)$ serão pequenos. A continuidade uniforme de F implica que, para qualquer A, $\sup_{x\in\mathbb{R}} |F(x+Ah) - F(x-Ah)| \to 0$ quando h tende para zero. Assim, $C_{2,n}$ será arbitrariamente pequeno se fizermos n tender para infinito. Atendendo à desigualdade (4.11) o teorema está demonstrado. □

Tal como no caso do estimador empírico, para que se possa estabelecer a convergência pontual do estimador da curva ROC pelo método do núcleo, é necessário demonstrar a convergência pontual do estimador da função quantil por este método. Repare-se que a demonstração do Teorema 4.1 é válida para qualquer estimador da função de distribuição que verifique o Teorema de Glivenko-Cantelli e que o próprio estimador seja uma função de distribuição. De facto, na demonstração, um dos passos utiliza a convergência do estimador. Por outro lado, a necessidade de aplicar a Proposição 2.2 requer a assumpção de que o estimador é uma função de distribuição.

Assim, a demonstração do resultado que apresentamos de seguida é omitida, visto ser análoga à demonstração do Teorema 4.1.

Teorema 4.12. *Seja $0 < 1 < p$, se $\xi_p = F^{-1}(p)$ for a única solução de $F(x-) \leq p \leq F(x)$, então*

$$\tilde{F}^{-1}(x) \xrightarrow{q.c.} F(x).$$

A partir deste resultado obtemos a convergência pontual do estimador $R\tilde{O}C$ da curva ROC e cuja a demonstração também será omitida, por ser análoga à do Teorema 4.5.

Teorema 4.13. *Seja* $0 < p < 1$. *Se* $\xi_{1-p} = F_0^{-1}(1-p)$ *for a única solução de* $F_0(x-) \leq 1-p \leq F_0(x)$ *e* F_1 *for contínua em* $F_0^{-1}(1-p)$, *então*

$$\tilde{\text{ROC}}(p) \xrightarrow{q.c.} \text{ROC}(p).$$

4.2.2. Convergência uniforme

A convergência uniforme do estimador da curva ROC pode ser facilmente deduzida com a ajuda do Teorema 4.10 apresentado anteriormente e com a ajuda de um segundo resultado demonstrado também em Horváth e Zhou [5] e que garante a convergência uniforme do estimador da função quantil:

Teorema 4.14. *Seja* F *uma função de distribuição com densidade* f *e sejam* $\tilde{F}$ *e* $\tilde{f} = \frac{1}{n}\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right)$ *os seus respectivos estimadores pelo método do núcleo. Suponhamos que o núcleo do estimador* $\tilde{f}$, K, *é suave, de variação limitada e cujo conjunto de descontinuidades tem medida de Lebesgue zero. Suponhamos também que,* $h \to 0$ *e* $\frac{nh}{\log(n)} \to \infty$ *quando* $n \to \infty$. *Defina-se* $t_F = \inf\{x : F(x) > 0\}$ *e* $t^F = \sup\{x : F(x) < 1\}$. *Suponhamos que* f *é uniformemente contínua em* $\mathbb{R}$ *e* $f(x) > 0$, $x \in]t_F, t^F[$. *Nestas condições, para todo o* $\epsilon > 0$, *temos*

$$\sup_{t\in[\epsilon,1-\epsilon]} \left|\tilde{F}^{-1}(t) - F^{-1}(t)\right| \xrightarrow{q.c.} 0$$

quando $n \to \infty$.

Demonstração. Esta demonstração faz uso da convergência uniforme do estimador da densidade nas condições do teorema, isto é:

$$\sup_{x\in\mathbb{R}} \left|\tilde{f}(x) - f(x)\right| \xrightarrow{q.c.} 0, \quad n \to \infty.$$

Esta propriedade foi demonstrada por Bertrand-Retali [1]. Deste resultado e tendo em conta que $f(x) > 0$, para $x \in]t_F, t^F[$, podemos assumir que

$$\inf_{A\leq x\leq B} \tilde{f}(x) > 0, \text{ se } n \geq n_0(\omega),$$

sendo $n_0(\omega)$ uma variável aleatória e onde $[A, B] \subset]t_F, t^F[$. Daqui vem que $\tilde{F}(x)$ é estritamente crescente em $[A, B]$ e assim,

$$\tilde{F}^{-1}\left(\tilde{F}(t)\right) = t, \text{ se } A = \tilde{F}^{-1}(a) \leq t \leq \tilde{F}^{-1}(b) = B.$$

Pelo Teorema 4.10 podemos assumir que $\epsilon > a$ e $1-\epsilon < b$, se $n \geq n_1(\omega)$, para alguma variável aleatória $n_1(\omega)$. Temos então $\tilde{F}\left(\tilde{F}^{-1}(t)\right) = t$, se $\epsilon \leq t \leq 1-\epsilon$ e portanto,

$$0 = \tilde{F}\left(\tilde{F}^{-1}(t)\right) - \tilde{F}\left(F^{-1}(t)\right) + \tilde{F}\left(F^{-1}(t)\right) - F\left(F^{-1}(t)\right). \tag{4.15}$$

Pelo Teorema 4.10, temos

$$\sup_{t\in[0,1]} \left|\tilde{F}\left(F^{-1}(t)\right) - F\left(F^{-1}(t)\right)\right| \xrightarrow{q.c.} 0, \tag{4.16}$$

e portanto, terá de se verificar que,

$$\sup_{t\in[\epsilon,1-\epsilon]} \left|\tilde{F}\left(\tilde{F}^{-1}(t)\right) - \tilde{F}\left(F^{-1}(t)\right)\right| \xrightarrow{q.c.} 0.$$

Pelo Teorema de valor médio, existe uma variável aleatória $\xi = \xi_t$ tal que

$$\sup_{t\in[\epsilon,1-\epsilon]} \tilde{f}(\xi(t)) \left|\tilde{F}^{-1}(t) - F^{-1}(t)\right| \geq \inf_{t\in[\epsilon,1-\epsilon]} \tilde{f}(\xi(t)) \sup_{t\in[\epsilon,1-\epsilon]} \left|\tilde{F}^{-1}(t) - F^{-1}(t)\right|.$$

Provámos que se $n \geq n_1(\omega)$,

$$\inf_{t\in[\epsilon,1-\epsilon]} \tilde{f}(\xi(t)) \geq \inf_{A\leq x\leq B} \tilde{f}(x).$$

Pela convergência do estimador da densidade, obtemos

$$\inf_{A\leq x\leq B} \tilde{f}(x) \longrightarrow \inf_{A\leq x\leq B} f(x) > 0.$$

Assim, de (4.15) e (4.16), conclui-se que

$$\sup_{t\in[\epsilon,1-\epsilon]} \left|\tilde{F}^{-1}(t) - F^{-1}(t)\right| \xrightarrow{q.c.} 0.$$

□

Com este resultado, facilmente se obtém a continuidade uniforme para o estimador da curva ROC. Os passos a seguir na demonstração são análogos à demonstração da convergência uniforme, pelo que a demonstração será omitida.

Teorema 4.17. *Seja F_1 uma função de distribuição contínua com densidade limitada e F_0 uma função de distribuição nas condições do Teorema 4.14. Para qualquer $\epsilon > 0$,*

$$\sup_{p\in[\epsilon,1-\epsilon]} \left|\tilde{\mathrm{ROC}}(p) - \mathrm{ROC}(p)\right| \xrightarrow{q.c.} 0.$$

4.2.3. A escolha da janela

A escolha da janela para o estimador do núcleo tem uma grande influência na estimação resultante. Essa escolha é feita consoante um determinado critério, segundo o qual a janela escolhida será assintoticamente óptima. Usualmente, o critério utilizado é o da minimização do erro quadrático médio integrado, o que no caso da curva ROC se traduziria na seguinte expressão:

$$EQMI = \int_{\mathscr{L}} E\left[\tilde{F}_1(\tilde{F}_0^{-1}(p)) - F_1(F_0^{-1}(p))\right]^2 dp,$$

para um conjunto $\mathscr{L} \subseteq [0,1]$. No entanto, se estivermos perante uma situação na qual as caudas da distribuição F_0 são muito mais leves do que as caudas da distribuição F_1, os erros cometidos por um estimador de F_0 nas suas caudas irão contribuir em grande parte para os erros cometidos pelo estimador correspondente de $F_1(F_0^{-1}(p))$, $0 < p < 1$. Por outras palavras, podemos ter uma estimação muito boa da curva ROC, excepto num pequeno intervalo em cada uma das extremidades, o que vai inflacionar o valor do erro quadrático médio, transmitindo uma ideia errada acerca do desempenho do estimador. Por essa razão, a minimização do erro quadrático médio integrado pode não ser um critério adequado para a escolha de uma janela óptima. A sugestão feita por Hall [4] é a de considerar outro o critério de optimalidade, dado pela seguinte expressão:

$$\alpha(\mathscr{L}) = \int_{\mathscr{L}} E\left[\tilde{F}_1(\tilde{F}_0^{-1}(p)) - F_1(F_0^{-1}(p))\right]^2 f_0\left(F_0^{-1}(1-p)\right) dp. \tag{4.18}$$

Para além disso, é possível provar que

$$\alpha(\mathscr{L}) \sim \beta(\mathscr{L}) \equiv \int_{F_0^{-1}(\mathscr{L})} \left\{E[\tilde{F}_0(x) - F_0(x)]^2 f_1^2(x) + E[\hat{F}_1(x) - F_1(x)]^2 f_0^2(x)\right\} dx$$

onde $F_0^{-1}(\mathscr{L})$ é o conjunto de pontos $F_0^{-1}(p)$ com $p \in \mathscr{L}$.

De facto, suponhamos que f_0 e f_1 têm derivadas contínuas e têm limite inferior maior que zero em $\mathscr{L}$. Defina-se $A = \tilde{F}_0 - F_0$, $B = \tilde{F}_1 - F_1$, $C = \tilde{F}_0^{-1} - F_0^{-1}$ e defina-se por I a função identidade. Da expansão de Taylor temos,

$$I = \tilde{F}_0(F_0^{-1} + C) = I + A(F_0^{-1}) + Cf_1(F_0^{-1}) + o(|A(f_0^{-1})| + |C|),$$

de onde segue que $C = -[A(F_0^{-1})/f(F_0^{-1})] + o(|A(F_0^{-1})|)$. Assim,

$$\tilde{F}_1(\tilde{F}_0^{-1}) - F_1(F_0^{-1}) = B(F_0^{-1}) - \frac{f_1(F_0^{-1})}{f_0(F_0^{-1})} A(F_0^{-1}) + o(|A(F_0^{-1})| + |B(F_0^{-1})|). \tag{4.19}$$

Repare-se na razão $f_1(F_0^{-1})/f_0(F_0^{-1})$. Visto que a variância de A é igual a $(1-F_0)F_0$, então o critério α_1 definido anteriormente é definido maioritariamente pelo valor de $(f_1/f_0)^2(1-F_0)F_0$ nas caudas, se esta quantidade não for limitada. No entanto, utilizando em vez disso o critério α, deduz-se a partir de (4.19) e da independência das amostras que

$$\alpha(\mathscr{L}) = [1+o(1)] \int_{F_0^{-1}(\mathscr{L})} [E(B^2)f_0^2 + E(A^2)f_1^2],$$

o que é equivalente a $\beta(\mathscr{L})$.

Peguemos no critério $\beta(\mathscr{L})$ e tomemos $F_0^{-1}(\mathscr{L})$ como sendo a recta real. Obtemos então o critério

$$\gamma(h_0, h_1) = \gamma_0(h_0) + \gamma_1(h_1),$$

onde

$$\gamma_0 = \int_{\mathbb{R}} E\left[\tilde{F}_0(x) - F_0(x)\right]^2 f_1^2(x)dx \quad \text{e} \quad \gamma_1 = \int_{\mathbb{R}} E\left[\tilde{F}_1(x) - F_1(x)\right]^2 f_0^2(x)dx. \tag{4.20}$$

O nosso objectivo seguinte é o de obter um desenvolvimento assintótico para γ_0 e γ_1, resultado esse que se apresenta de seguida.

Teorema 4.21. *Suponhamos que K é uma densidade de probabilidade simétrica. Suponhamos também que F_0 e F_1 admitem derivada de segunda ordem limitada e contínua em $\mathbb{R}$ e que f_0 e f_1 são de quadrado integrável. Nestas condições, γ_0 admite o seguinte desenvolvimento assintótico:*

$$\gamma_0 = -A_0 h_0 + B_0 h_0^4 + C_0 + o\left(\frac{h_0}{n_0} + h_0^4\right)$$

onde

$$A_0 = \frac{1}{n_0} \int L(z)\,(1-L(z))\,dz \int f_0(x) f_1(x)^2 dx,$$

$$B_0 = \frac{1}{4}\left(\int u^2 K(u)du\right)^2 \int f_0'(x)^2 f_1(x)^2 dx,$$

$$C_0 = \frac{1}{n_0} \int F_0(x)\,(1-F_0(x))\,f_1(x)^2 dx.$$

Por seu lado, o desenvolvimento assintótico de γ_1 é dado por:

$$\gamma_1 = -A_1 h_1 + B_1 h_1^4 + C_1 + o\left(\frac{h_1}{n_1} + h_1^4\right)$$

onde

$$A_1 = \frac{1}{n_1}\int L(z)\,(1-L(z))\,dz \int f_1(x) f_0(x)^2 dx,$$

$$B_1 = \frac{1}{4}\left(\int u^2 K(u) du\right)^2 \int f_1'(x)^2 f_0(x)^2 dx,$$

$$C_1 = \frac{1}{n_1}\int F_1(x)\,(1-F_1(x))\,f_0(x)^2 dx.$$

Demonstração. Analisemos em primeiro lugar o caso de γ_0, mais especificamente na parcela $E\left[\tilde{F}_0(x) - F_0(x)\right]^2$. Temos que,

$$E\left[\tilde{F}_0(x) - F_0(x)\right]^2 = \mathrm{Var}\left(\tilde{F}_0(x)\right) + \left(E[\tilde{F}_0(x) - F_0(x)]\right)^2$$

e assim,

$$\gamma_0 = \int \mathrm{Var}\left(\tilde{F}_0(x)\right) f_1^2(x) dx + \int \left(E[\tilde{F}_0(x) - F_0(x)]\right)^2 f_1^2(x) dx.$$

Comecemos por calcular a esperança de $\hat{F}_0(x)$. Utilizando o facto de F_0 ser diferenciável duas vezes e integrando por partes, temos:

$$\begin{aligned} E\left[\tilde{F}_0(x)\right] &= \int L\left(\frac{x-y}{h_0}\right) dF_0(y) \\ &= F_0(x) + h_0 f_0(x) \int uK(u)du + h_0^2 \int u^2 K(u) \int_0^1 (1-s) f_0'(x - suh_0) ds du. \end{aligned}$$

Como K é simétrica, temos $\int uK(u)du = 0$ e assim

$$\begin{aligned} \int \left(E[\tilde{F}_0(x) - F_0(x)]\right)^2 f_1^2(x) dx &= h_0^4 \int \left[\int u^2 K(u) \int_0^1 (1-s) f_0'(x - suh_0) ds du\right]^2 f_1^2(x) dx \\ &= h_0^4 \iiint \int_0^1 \int_0^1 u^2 K(u) v^2 K(v) (1-s)(1-t) \\ &\quad \times f_0'(x - suh_0) f_0'(x - tvh_0) f_1^2(x) du dv ds dt dx \end{aligned} \tag{4.22}$$

As condições do teorema dizem-nos que F_0 é contínua e portanto, $\left(E[\tilde{F}_0(x) - F_0(x)]\right)^2 f_1^2(x)$ converge pontualmente para 0, quando $n_0 \to \infty$. Majorando a parcela 4.22 por

$$g(x) = u^2 |K(u)| v^2 |K(v)| (1-s)(1-t) \left(\sup_{x\in\mathbb{R}} |f_0'(x)|\right)^2 f_1^2(x),$$

e sendo $g(x)$ integrável, estamos nas condições do Teorema de Convergência Dominada de Lebesgue (ver Serfling [9] (p.11)), o que nos permite concluir que

$$h_0^4 \iiint \int_0^1 \int_0^1 u^2 K(u) v^2 K(v) (1-s)(1-t) f_0'(x - suh_0) f_0'(x - tvh_0) f_1^2 du dv ds dt dx$$

também é integrável. Para além disso, temos também $\int u^2|K(u)|du < \infty$ e daqui obtemos que

$$\int \left(E[\tilde{F}_0(x) - F_0(x)]\right)^2 f_1^2(x)dx = h_0^4 \left[\left(\int u^2 K(u)du\right)^2 \frac{1}{4}\int f_0'(u)^2 f_1(u)du + o(1)\right]. \tag{4.23}$$

Calculemos agora a variância:

$$\begin{aligned}
\text{Var}\left(\tilde{F}_0(x)\right) &= \text{Var}\left(\frac{1}{n_0}\sum_{i=1}^{n_0} L\left(\frac{x - X_{0i}}{h_0}\right)\right) \\
&= \frac{1}{n_0}\text{Var}\left(L\left(\frac{x - X_0}{h_0}\right)\right) \\
&= \frac{1}{n_0}\left\{E\left(L\left(\frac{x - X_0}{h_0}\right)^2\right) - E\left(L\left(\frac{x - X_0}{h_0}\right)\right)^2\right\} \\
&= \frac{1}{n_0}\left\{\int L\left(\frac{x - y}{h_0}\right)^2 f_0(y)dy - \left(E[\tilde{F}_0(x)]\right)^2\right\} \\
&= \frac{1}{n_0}\int L\left(\frac{x - y}{h_0}\right)^2 f_0(y)dy - \frac{1}{n_0}\left(E[\tilde{F}_0(x) - F_0(x) + F_0(x)]\right)^2 \\
&= \frac{1}{n_0}\int L\left(\frac{x - y}{h_0}\right)^2 f_0(y)dy \\
&\quad - \frac{1}{n_0}\left\{\left(E[\tilde{F}_0(x) - F_0(x)]\right)^2 - 2E[\tilde{F}_0(x) - F_0(x)]F_0(x) + F_0(x)^2\right\}
\end{aligned}$$

Analisemos por partes a segunda parcela. Aplicando o que deduzimos à definição de γ_0, de (4.23), temos que

$$\int \left(E[\tilde{F}_0(x) - F_0(x)]\right)^2 f_1(x)^2 du = O(h_0^4) \tag{4.24}$$

Temos também que,

$$\int F_0(x)^2 f_1(x)^2 dx < \infty \tag{4.25}$$

Resta ver o que acontece para $\int E[\hat{F}_0(x) - F_0(x)]F_0(x)f_1(x)dx$. Façamos a seguinte dedução

$$\begin{aligned}
\int E[\tilde{F}_0(x) - F_0(x)]F_0(x)f_1(x)^2 dx &\leq \left|\int E[\tilde{F}_0(x) - F_0(x)]f_1(x)F_0(x)f_1(x)dx\right| \\
&\leq \left(\int \left(E[\tilde{F}_0(x) - F_0(x)]\right)^2 f_1(x)^2 dx\right)^{\frac{1}{2}} \\
&\quad \times \left(\int F_0(x)^2 f_1(x)^2 dx\right)^{\frac{1}{2}}.
\end{aligned}$$

Mais uma vez, do resultado (4.23), conclui-se que

$$\int E[\hat{F}_0(x) - F_0(x)]F_0(x)f_1(x)dx = O(h^2) \tag{4.26}$$

Vejamos agora o que acontece para $\frac{1}{n_0}\iint L\left(\frac{x-y}{h_0}\right)^2 f_0(y)dyf_1(x)^2$, começando por calcular o que temos no interior do primeiro integral:

$$\begin{aligned}
\int L\left(\frac{x-y}{h_0}\right)^2 f_0(y)dy &= F_0(y)L\left(\frac{x-y}{h_0}\right)^2\Big]_{-\infty}^{+\infty} \\
&\quad - \int F_0(y)2L\left(\frac{x-y}{h_0}\right)K\left(\frac{x-y}{h_0}\right)\left(-\frac{1}{h_0}\right)dy \\
&= \frac{1}{h_0}\int F_0(y)B\left(\frac{x-y}{h_0}\right)dy \\
&= \int B(z)F_0(x-zh_0)dz \\
&= \int B(z)\Big\{F_0(x) - zh_0f_0(x) + z^2h_0^2\int_0^1(1-t) \\
&\quad \times f_0'(x-tzh_0)dt\Big\}dz \\
&= F_0(x)\int B(z)dz - h_0f_0(x)\int zB(z)dz \\
&\quad + h_0^2\int z^2B(z)\int_0^1(1-t)f_0'(x-tzh_0)dtdz,
\end{aligned}$$

onde $B(u) = 2L(u)K(u)$. Observe-se que

$$\int B(z)dz = 2\int L(z)K(z) = 2\int L'(z)L(z)dz = 2\,\frac{L(z)^2}{2}\Big]_{-\infty}^{+\infty} = 1.$$

Assim,

$$\begin{aligned}
\int \mathrm{Var}\left(\tilde{F}_0(x)\right)f_1(x)^2dx &= \frac{1}{n_0}\int\Big[F_0(x) - h_0f_0(x)\int zB(z)dz \\
&\quad + h_0^2\int z^2B(z)\int_0^1(1-t)f_0'(x-tzh_0)dtdz\Big]f_1(x)^2dx \\
&\quad - \frac{1}{n_0}\int F_0(x)^2f_1(x)^2dx + O\left(\frac{h^2}{n_0}\right) \\
&= \frac{1}{n_0}\Big[\int F_0(x)f_1(x)^2dx - h_0\int zB(z)dz\int f_0(x)f_1(x)^2dx \\
&\quad + h_0^2\iiint_0^1 z^2B(z)(1-t)f_0'(x-tzh_0)f_1(x)^2dzdtdx\Big] \\
&\quad - \frac{1}{n_0}\int F_0(x)^2f_1(x)^2dx + O\left(\frac{h^2}{n_0}\right)
\end{aligned}$$

Peguemos na última parcela calculada. Repare-se que

$$z^2B(z)(1-t)f_0'(x-tzh_0)f_1(x)^2 \le z^2\left|B(z)\right|(1-t)\mathrm{sup}_{x\in\mathbb{R}}\left|f_0'(x)\right|f_1(x)^2.$$

Para além disso, da continuidade de F_0, sabemos que $\mathrm{Var}\left(\tilde{F}_0(x)\right)f_1(x)^2$ converge pontualmente para $\mathrm{Var}\left(F_0(x)\right)f_1(x)^2$, quando $n_0 \to \infty$. Estamos nas condições de

aplicar o Teorema da Convergência Dominada de Lebesgue, concluindo que o integral

$$\iiint_0^1 z^2 B(z)(1-t) f_0'(x - tzh_0) f_1(x)^2 dz dt dx$$

existe e é finito, obtendo-se assim um desenvolvimento assintótico para a variância:

$$\begin{aligned}\int \mathrm{Var}\left(\tilde{F}_0(x)\right) f_1(x)^2 dx = \frac{1}{n_0}\int F_0(x)\left(1 - F_0(x)\right) f_1(x)^2 dx \\ - \frac{h}{n_0}\int zB(z)dz \int f_0(x) f_1(x)^2 dx + O\left(\frac{h^2}{n_0}\right). \quad (4.27)\end{aligned}$$

Juntando os resultados (4.23) e (4.27), obtemos já um desenvolvimento assintótico para γ_0:

$$\begin{aligned}\gamma_0 &= \int E\left[\tilde{F}_0(x) - F_0(x)\right]^2 f_1(x)^2 dx \\ &= \frac{1}{n_0}\int F_0(x)\left(1 - F_0(x)\right) f_1(x)^2 dx - \frac{h_0}{n_0}\int zB(z)dz \int f_0(x) f_1(x)^2 dx \\ &\quad + \frac{h_0^4}{4}\left(\int u^2 K(u) du\right)^2 \int f_0'(x)^2 f_1(x)^2 dx + o\left(\frac{h_0}{n_0}\right) + o(h_0^4). \quad (4.28)\end{aligned}$$

Concentremo-nos na parcela $\int zB(z)dz$. Para podermos efectuar a minimização, é necessário conhecer o seu sinal em todo $\mathbb{R}$. Sejam a e b dois reais tais que $a < b$. Suponhamos que K é uma densidade tal que:

$$\int xK(x)dx = 0 \quad \text{e} \quad \int x^2 K(x)dx < +\infty$$

Sabemos que,

$$\begin{aligned}\int_a^b zB(z)dz &= \int_a^b z2K(z)L(z)dz \\ &= \int_a^b z\left(L(z)^2\right)' dz \\ &= zL(z)^2\big]_a^b - \int_a^b L(z)^2 dz \\ &= bL(b)^2 - aL(a)^2 - \int_a^b L(z)^2 dz.\end{aligned}$$

Para prosseguirmos a nossa dedução, vamos fazer uso de um pequeno artifício. Repare-se que,

$$\begin{aligned}\int_a^b zK(z)dz &= zL(z)\big]_a^b - \int_a^b L(z)dz \\ &= bL(b) - aL(a) - \int_a^b L(z)dz\end{aligned}$$

e portanto, $\int_a^b zK(z)dz - bL(b) + aL(a) + \int_a^b L(z)dz = 0$. Assim,

$$\begin{aligned}\int_a^b zB(z)dz &= \int_a^b zB(z)dz + \int_a^b zK(z)dz - bL(b) + aL(a) + \int_a^b L(z)dz \\ &= bL(b)\left(L(b)-1\right) - aL(a)\left(L(a)-1\right) \\ &\quad - \int_a^b zK(z)dz + \int_a^b L(z)\left(1-L(z)\right)dz.\end{aligned}$$

Se fizermos $a = -\infty$ e $b = +\infty$ então

$$\int zB(z)dz = \int L(z)\left(1-L(z)\right)dz > 0.$$

Aplicando esta última dedução a (4.28), chegamos ao resultado pretendido para o desenvolvimento assintótico de γ_0. A demonstração para o caso de γ_1 é feita de forma análoga, pelo que será omitida. □

A partir das expressões que obtivemos para o desenvolvimento assintótico de γ_0 e γ_1, facilmente conseguimos calcular as janelas h_0 e h_1 assintoticamente óptimas de γ_0 e γ_1, respectivamente:

$$h_0 = \left(\frac{\int L(z)\left(1-L(z)\right)dz \int f_0(x) f_1(x)^2 dx}{n_0 \left(\int u^2 K(u)du\right)^2 \int f_0'(x)^2 f_1(x)^2 dx} \right)^{\frac{1}{3}}$$

e

$$h_1 = \left(\frac{\int L(z)\left(1-L(z)\right)dz \int f_1(x) f_0(x)^2 dx}{n_1 \left(\int u^2 K(u)du\right)^2 \int f_1'(x)^2 f_0(x)^2 dx} \right)^{\frac{1}{3}}.$$

Como se pode observar, estas expressões não podem ser usadas directamente no momento da estimação, pois dependem de quantidades que são desconhecidas. Uma possível solução para cotornar este problema, será o de adoptar um modelo de referência para o cálculo das janelas. Este método tem as suas raízes no método das distribuições de referência para a escolha da janela no estimador do núcleo da densidade. No nosso caso, iremos utilizar o modelo binormal como modelo de referência, o que, consequentemente, leva a que as janelas sejam calculadas tomando para f_0 e f_1 densidades normais. Obtemos então

$$\begin{aligned}h_0 = n_0^{-1/3} \Bigg(& \frac{4\sqrt{\pi}\int L(u)\left(1-L(u)\right)du}{\int u^2 K(u)du} \frac{\sigma_0^3(\sigma_0^2+\sigma_1^2)^{5/2}}{(\sigma_1^2+2\sigma_0^2)^{1/2}[\sigma_1^4+\sigma_1^2\sigma_0^2+2\sigma_0^2(\mu_0-\mu_1)^2]} \\ & \times \exp\left[\frac{(\mu_0-\mu_1)^2\sigma_0^2}{(\sigma_0^2+\sigma_1^2)(2\sigma_0^2+\sigma_1^2)}\right]\Bigg)^{1/3}\end{aligned}$$

e

$$h_1 = n_1^{-1/3} \left(\frac{4\sqrt{\pi} \int L(u)\,(1 - L(u))\,du}{\int u^2 K(u) du} \frac{\sigma_1^3 (\sigma_1^2 + \sigma_0^2)^{5/2}}{(\sigma_0^2 + 2\sigma_1^2)^{1/2} [\sigma_0^4 + \sigma_0^2 \sigma_1^2 + 2\sigma_1^2 (\mu_1 - \mu_0)^2]} \right.$$
$$\left. \times \exp \left[\frac{(\mu_1 - \mu_0)^2 \sigma_1^2}{(\sigma_1^2 + \sigma_0^2)(2\sigma_1^2 + \sigma_0^2)} \right] \right)^{1/3},$$

onde μ_0 e μ_1, σ_0^2 e σ_1^2 representam a média e a variância das respectivas populações. Na prática, estes valores serão substituídos pelas médias e variâncias amostrais de cada uma das populações. Tal como no caso da estimação da densidade, será de esperar que este método funcione bem desde que as populações consideradas não tenham distribuições muito distantes da distribuição normal.

Capítulo 5

Estudo de simulação

Os capítulos anteriores incidiram sobre as noções teóricas da curva ROC e sobre as propriedades assintóticas dos diversos estimadores considerados. Neste capítulo, pretendemos estudar o comportamento dos estimadores a distância finita. Para esse efeito, iremos realizar diversas simulações de situações práticas, com o intuito de comparar o desempenho dos métodos de estimação que possuimos para a aproximação da curva ROC, a partir de uma determinada amostra. O desempenho de cada um dos estimadores irá ser avaliado através da diferença entre a área sob a curva verdadeira e a área sob a curva estimada e de uma medida de erro, que neste caso será baseada na medida do critério de optimalidade visto no capítulo anterior, para a escolha das janelas do estimador do núcleo (ver (4.18)), à qual chamaremos *erro quadrático integrado ponderado* e será dada por:

$$\text{EQIP} = \int_0^1 \left(\text{R}\check{\text{O}}\text{C}(p) - \text{ROC}(p)\right)^2 f_0\left(F_0^{-1}(1-p)\right) dp,$$

onde $\text{R}\check{\text{O}}\text{C}(p)$ representa um estimador da curva ROC.

As simulações vão ser realizadas com a ajuda do software informático *R* [8], uma ferramenta muito útil, desenvolvida precisamente para a computação estatística. As simulações irão todas seguir a mesma estrutura; em primeiro lugar, irá ser calculada a verdadeira curva ROC e a sua respectiva área sob a curva. Seguidamente, o programa irá gerar duas amostras, uma de indíviduos saudáveis e outra de indivíduos doentes, ambas de tamanho 200, segundo as distribuições que foram atríbuidas a cada uma das populações no caso em simulação. O próximo passo é o da estimação, onde vão ser utilizados os quatro métodos que foram estudados até agora: estimação paramétrica, estimação paramétrica com transformação de Box-Cox (se necessária) e estimação não-paramétrica, empírica e pelo método do núcleo. Refira-se que neste último caso, o núcleo utilizado na estimação da curva será o núcleo de Epanechnikov, definido por: $K(x) = \frac{3}{4}(1-x)^2\mathbb{I}_{[-1,1]}(x)$. Após serem obtidas as estimações, proceder-se-á ao cálculo do erro e da área sob as curvas estimadas. Este processo irá ser repetido mil vezes. No final, apresentaremos os boxplots onde estarão repre-

sentados os resultados dessas repetições, para a comparação dos diferentes métodos de estimação. Apresentaremos também tabelas com as médias e desvios padrão das medidas calculadas.

5.1. Populações consideradas

No nosso estudo irão ser realizadas seis simulações diferentes, sendo as amostras geradas segundo as seguintes distribuições, cujas densidades são apresentadas na Figura 5.1 e na Figura 5.2 em conjunto com as suas respectivas curvas ROC.

- **Cenário 1**: $X_0 \sim N(4.5, 0.09)$ e $X_1 \sim N(5.5, 0.25)$;
- **Cenário 2**: $X_0 \sim SN(4.5, 0.09, 15)$ e $X_1 \sim SN(5.5, 0.25, 15)$, onde $SN(\mu, \sigma, \alpha)$ representa a lei normal assimétrica e α é o parâmetro de forma. A densidade desta lei é dada por $f(x; \mu, \sigma, \alpha) = \frac{2}{\sigma}\phi\left(\frac{x-\mu}{\sigma}\right)\Phi\left(\alpha\left(\frac{x-\mu}{\sigma}\right)\right)$, onde ϕ e Φ representam a densidade e a distribuição normal centrada e reduzida;
- **Cenário 3**: $X_0 \sim LN(0, 0.5)$ e $X_1 \sim LN(1, 0.5)$, onde $LN(\log -\mu, \log -\sigma)$ representa a lei log-normal, $\log -\mu$ e $\log -\sigma$ representam a média e o desvio padrão na escala logarítmica;
- **Cenário 4**: $X_0 \sim \Gamma(2, 2)$ e $X_1 \sim \Gamma(3, 1)$, onde $\Gamma(\alpha, \frac{1}{\theta})$ representa a distribuição gamma, α o parâmetro de forma e θ o parâmetro de escala;
- **Cenário 5**: $X_0 \sim N(5, 1)$ e X_1 segue mistura de duas distribuições normais tal que, $X_1 \sim NorMix(\mu, \sigma)$, onde $\mu = (6, 8)$ é o vector das médias e $\sigma = (0.75, 1)$ é o vector das variâncias da mistura;
- **Cenário 6**: $X_0 \sim NorMix(\mu_0, \sigma_0)$, onde $\mu_0 = (5, 8)$ e $\sigma_0 = (1, 1)$ e $X_1 \sim NorMix(\mu_1, \sigma_1)$, onde $\mu_1 = (9, 13)$ e $\sigma_1 = (1, 3)$.

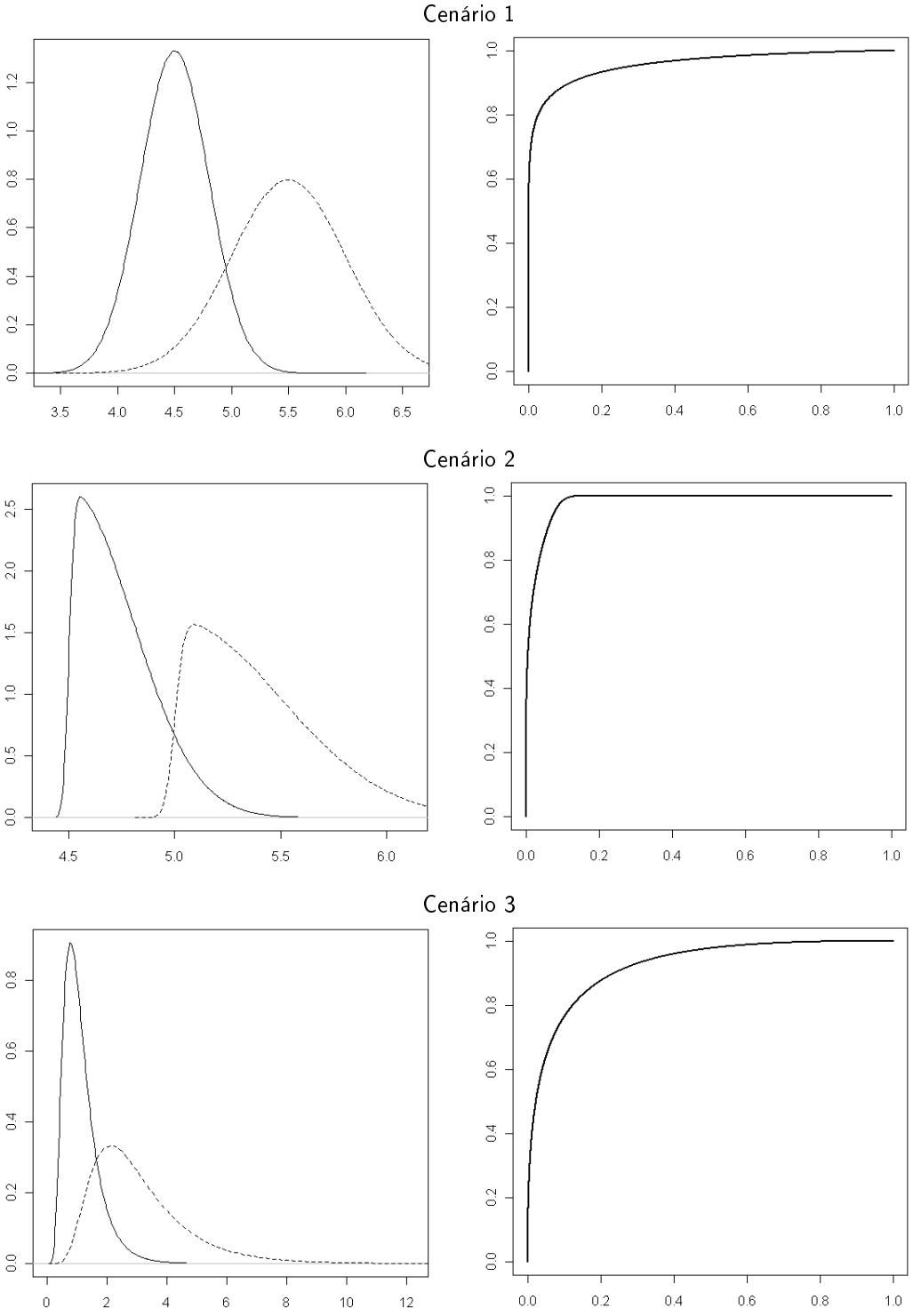


Figura 5.1: *Esquerda: Densidades de X_0 (linha) e X_1 (tracejado). Direita: Curva ROC*

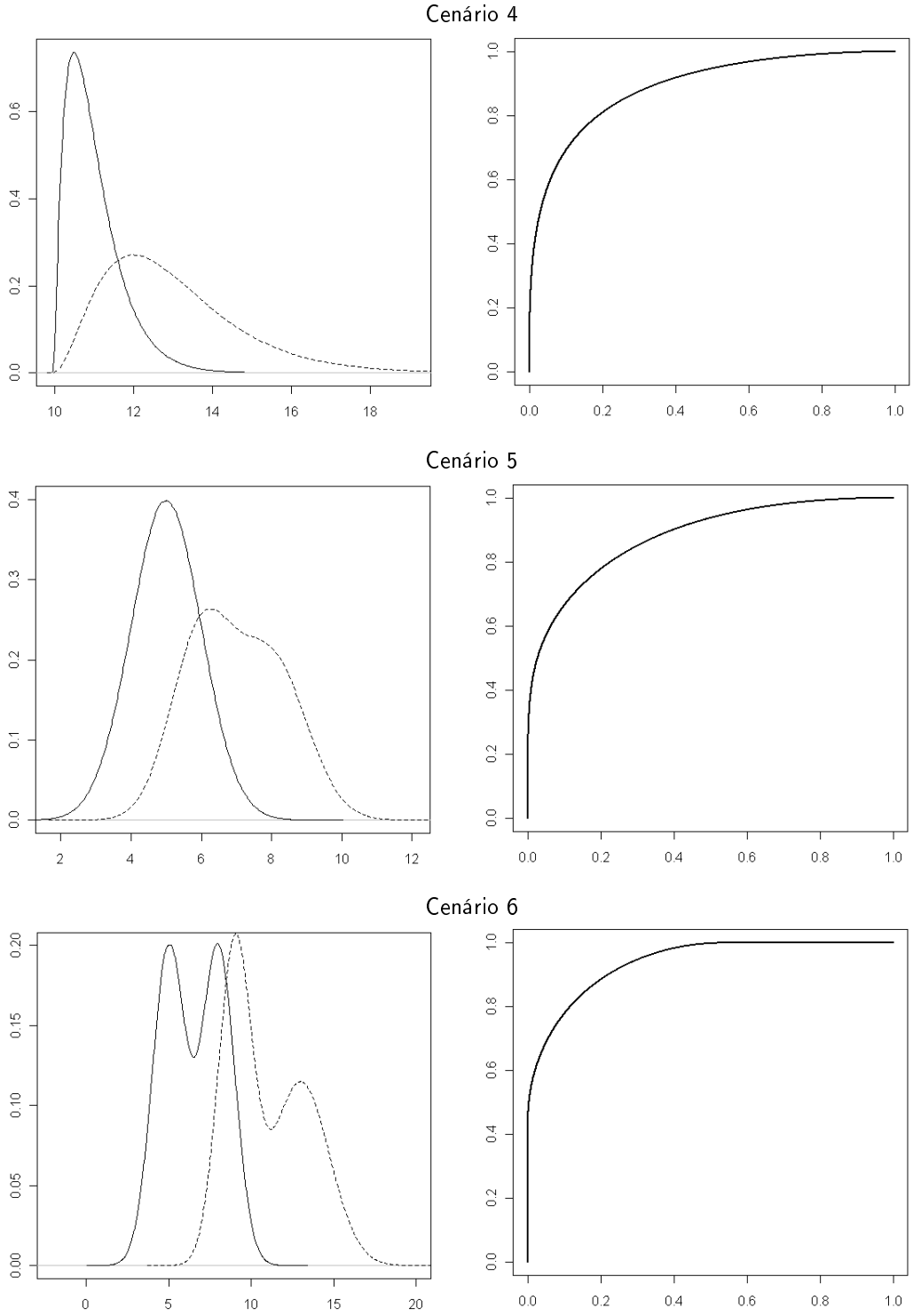


Figura 5.2: *Esquerda: Densidades de X_0 (linha) e X_1 (tracejado). Direita: Curva ROC*

5.2. Resultados

5.2.1. Cenário 1

	AUC-AŬC		EQIP	
Métodos de Estimação	**Média**	**Desvio-padrão**	**Média**	**Desvio-padrão**
Paramétrico	8.74×10^{-4}	9.15×10^{-3}	9.70×10^{-5}	1.42×10^{-4}
Empírico	5.91×10^{-4}	9.97×10^{-3}	1.81×10^{-4}	1.81×10^{-4}
Núcleo	7.13×10^{-4}	9.98×10^{-3}	1.80×10^{-4}	1.85×10^{-4}

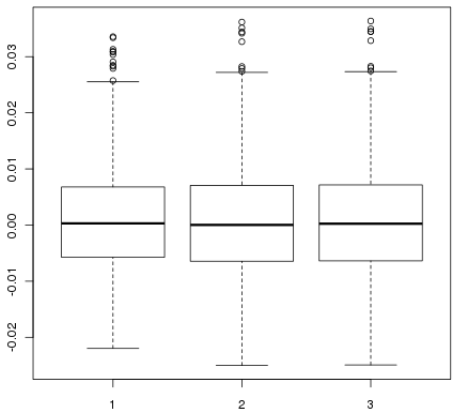


Figura 5.3: Valores de AUC: 1-Estimador Paramétrica, 2-Estimador Empírica, 3-Estimador do Núcleo

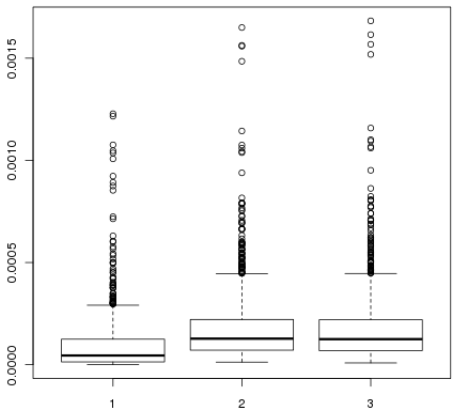


Figura 5.4: Valores do erro: 1-Estimador Paramétrica, 2-Estimador Empírica, 3-Estimador do Núcleo

5.2.2. Cenário 2

	AUC-AŬC		EQIP	
Métodos de Estimação	**Média**	**Desvio-padrão**	**Média**	**Desvio-padrão**
Paramétrico	1.25×10^{-2}	6.31×10^{-3}	9.94×10^{-4}	3.79×10^{-4}
Box-Cox	-5.74×10^{-5}	4.91×10^{-3}	1.05×10^{-4}	7.42×10^{-5}
Empírico	-4.47×10^{-6}	5.45×10^{-3}	1.18×10^{-4}	1.27×10^{-4}
Núcleo	9.06×10^{-5}	5.47×10^{-3}	1.26×10^{-4}	1.45×10^{-4}

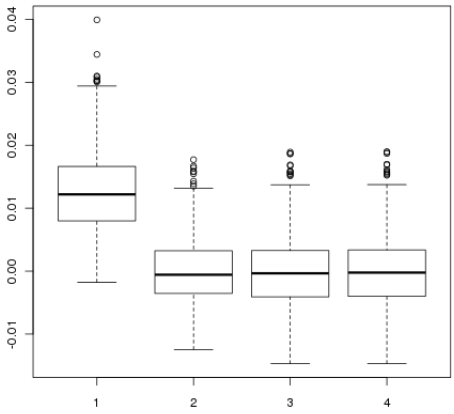


Figura 5.5: Valores de AUC: 1-Estimador Paramétrica, 2-Estimador Box-Cox, 3-Estimador Empírica, 4-Estimador do Núcleo.

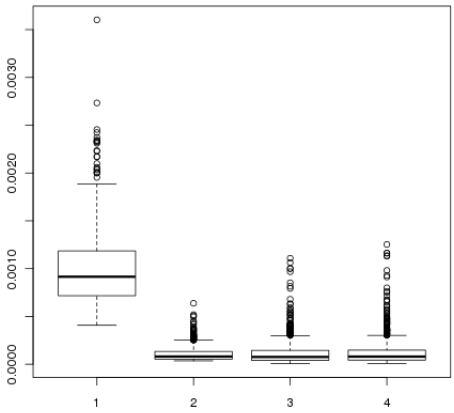


Figura 5.6: Valores do erro: 1-Estimador Paramétrica, 2-Estimador Box-Cox, 3-Estimador Empírica, 4-Estimador do Núcleo.

5.2.3. Cenário 3

	AUC-AŬC		EQIP	
Métodos de Estimação	**Média**	**Desvio-padrão**	**Média**	**Desvio-padrão**
Paramétrico	5.32×10^{-2}	1.68×10^{-2}	1.16×10^{-3}	2.35×10^{-4}
Box-Cox	6.31×10^{-4}	1.23×10^{-2}	1.31×10^{-4}	1.67×10^{-4}
Empírico	4.16×10^{-4}	1.28×10^{-2}	2.14×10^{-4}	1.80×10^{-4}
Núcleo	2.93×10^{-3}	1.33×10^{-2}	2.01×10^{-4}	1.85×10^{-4}

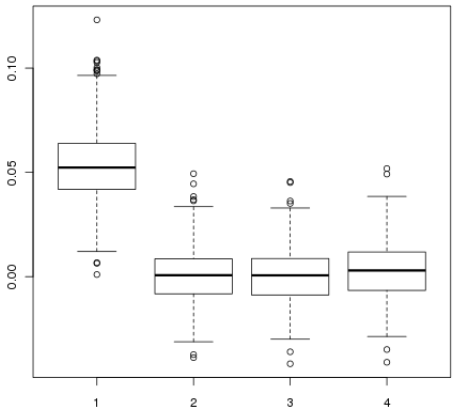


Figura 5.7: Valores de AUC: 1-Estimador Paramétrica, 2-Estimador Box-Cox, 3-Estimador Empírica, 4-Estimador do Núcleo

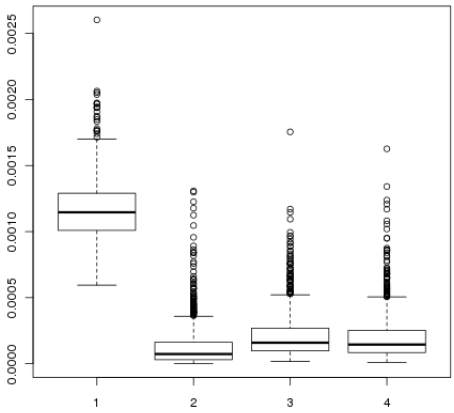


Figura 5.8: Valores do erro: 1-Estimador Paramétrica, 2-Estimador Box-Cox, 3-Estimador Empírica, 4-Estimador do Núcleo

5.2.4. Cenário 4

	AUC-AŬC		EQIP	
Métodos de Estimação	**Média**	**Desvio-padrão**	**Média**	**Desvio-padrão**
Paramétrico	3.06×10^{-2}	1.55×10^{-2}	1.88×10^{-3}	7.17×10^{-4}
Box-Cox	-1.10×10^{-3}	1.53×10^{-2}	1.21×10^{-4}	1.44×10^{-4}
Empírico	-3.85×10^{-4}	1.57×10^{-2}	1.90×10^{-4}	1.60×10^{-4}
Núcleo	3.06×10^{-3}	1.62×10^{-2}	1.70×10^{-4}	1.70×10^{-4}

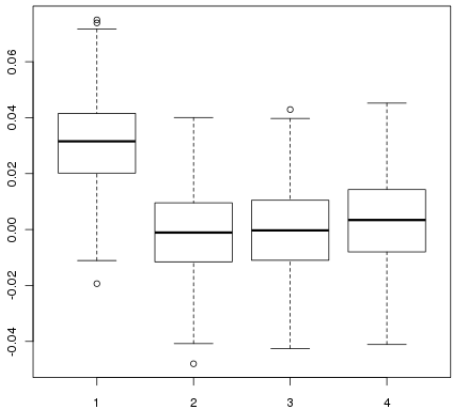


Figura 5.9: Valores de AUC: 1-Estimador Paramétrica, 2-Estimador Box-Cox, 3-Estimador Empírica, 4-Estimador do Núcleo

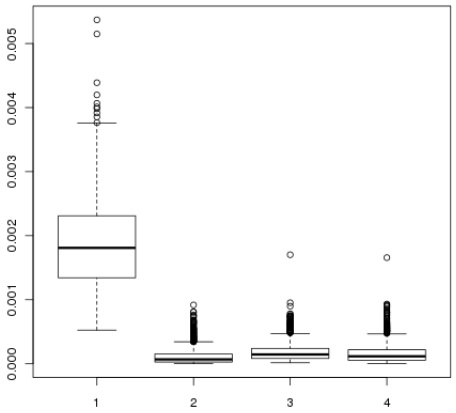


Figura 5.10: Valores do erro: 1-Estimador Paramétrica, 2-Estimador Box-Cox, 3-Estimador Empírica, 4-Estimador do Núcleo

5.2.5. Cenário 5

	AUC-AŬC		EQIP	
Métodos de Estimação	**Média**	**Desvio-padrão**	**Média**	**Desvio-padrão**
Paramétrico	-1.49×10^{-3}	1.54×10^{-2}	1.51×10^{-4}	1.35×10^{-4}
Box-Cox	-2.13×10^{-3}	1.54×10^{-2}	1.39×10^{-4}	1.40×10^{-4}
Empírico	-2.69×10^{-6}	1.65×10^{-2}	1.77×10^{-4}	1.72×10^{-4}
Núcleo	3.49×10^{-3}	1.67×10^{-2}	1.56×10^{-4}	1.80×10^{-4}

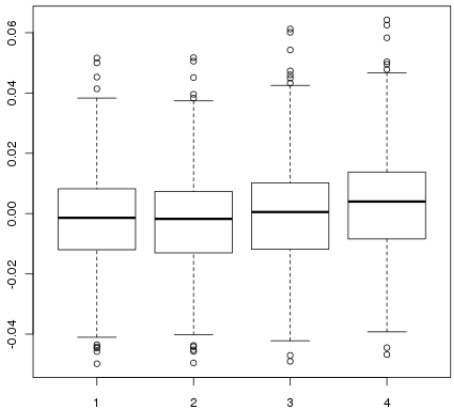


Figura 5.11: Valores de AUC: 1-Estimador Paramétrica, 2-Estimador Box-Cox, 3-Estimador Empírica, 4-Estimador do Núcleo

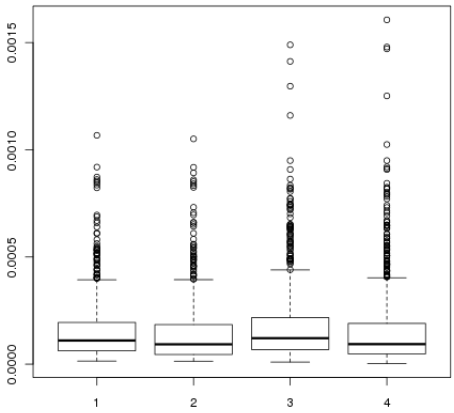


Figura 5.12: Valores do erro: 1-Estimador Paramétrica, 2-Estimador Box-Cox, 3-Estimador Empírica, 4-Estimador do Núcleo

5.2.6. Cenário 6

	AUC-AŬC		EQIP	
Métodos de Estimação	**Média**	**Desvio-padrão**	**Média**	**Desvio-padrão**
Paramétrico	8.11×10^{-3}	9.44×10^{-3}	9.19×10^{-5}	3.45×10^{-5}
Box-Cox	5.51×10^{-3}	9.27×10^{-3}	5.83×10^{-5}	3.66×10^{-5}
Empírico	5.59×10^{-4}	1.06×10^{-2}	6.25×10^{-5}	6.08×10^{-5}
Núcleo	2.38×10^{-2}	1.84×10^{-2}	2.14×10^{-4}	1.84×10^{-4}

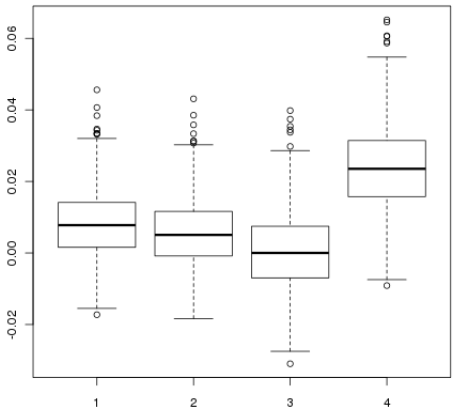


Figura 5.13: Valores de AUC: 1-Estimador Paramétrica, 2-Estimador Box-Cox, 3-Estimador Empírica, 4-Estimador do Núcleo

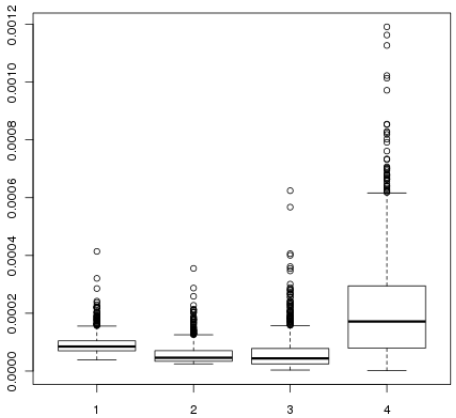


Figura 5.14: Valores do erro: 1-Estimador Paramétrica, 2-Estimador Box-Cox, 3-Estimador Empírica, 4-Estimador do Núcleo

5.3. Discussão dos resultados e conclusão

Observando os resultados que foram obtidos, constatamos que, como seria de esperar, o desempenho do estimador paramétrico sem transformação vai piorando à medida que as distribuições de cada uma das populações se afasta da distribuição normal, o que faz todo o sentido, pois na prática é imposto o modelo binormal a dados que podem não se adequar a esse modelo. Ainda assim, este estimador não se comporta muito mal quando se está na presença de misturas de distribuições normais.

Apesar da janela óptima do estimador não-paramétrico pelo método do núcleo ser obtida tendo como referência distribuições normais, este estimador tem na maioria dos casos um bom desempenho, sendo o seu pior desempenho registado no último cenário. De facto, o desempenho do estimador não é muito afectado nos cenários de distribuições não normais, sendo a quebra de desempenho apenas notável no cenário das distribuições bimodais.

Os métodos de estimação paramétrica com transformação de Box-Cox e os métodos empíricos revelam-se como sendo os mais eficazes na estimação da curva ROC. De facto, podemos observar que ambos proporcionam bons resultados em todos os cenários feitos neste estudo, tanto no cálculo da área sob a curva ROC como no erro da estimação.

Concluindo, a estimação paramétrica, pela sua simplicidade, é uma boa solução, caso consigamos garantir a normalidade dos dados. Como vimos, o estimador empírico tem um bom desempenho e não necessita de assumpções acerca das distribuições dos dados e é particularmente útil para o caso de apenas necessitarmos de estimação o valor da área sob a curva, visto que a estatística de Mann-Whitney é realtivamente simples de calcular a partir da amostra. O facto de não devolver estimações suaves da curva ROC e de poder falhar no caso das amostras à disposição serem demasiado pequenas, são os dois notáveis inconvenientes da estimação empírica. No caso do estimador do núcleo, apesar de este permitir estimar uma curva ROC suave, este método de estimação é também o que tem a implementação para a prática mais complexa, sobretudo tendo em conta o problema da escolha das janelas. Assim, a estimação paramétrica com transformação de Box-Cox parece ser a escolha acertada para a estimação da curva ROC em casos práticos, pois combina a simplicidade da estimação paramétrica com a vantagem de devolver uma estimação suave da curva ROC, o que é possibilitado pelo bom ajuste dos dados à normalidade proporcionado pela transformação de Box-Cox.

Bibliografia

[1] **Bertrand-Retali, M.** (1978), "Convergence uniforme d'un estimateur de la densité par la méthode du noyau.", *Rev. Roumaine Math. Pures Appl. 13*, p. 361-385.

[2] **Box, G. E. P., Cox, D. R.** (1964), "An Analysis of Transformations", *Journal of the Royal Statistical Society. Series B (Methodological), Vol. 26, No. 2.*, p. 211-252.

[3] **Gonçalves, E. , Lopes, N. M.** (2003), *Estatística, Teoria Matemática e Aplicações*. Escolar Editora, Lisboa.

[4] **Hall, P.G., Hyndman R. J.** (2003), "Improved methods for bandwidth selection when estimating ROC curves", *Statistics & Probability Letters 64*, p. 181-189.

[5] **Horváth, L., Horváth, Z., Zhou W.** (2008), "Confidence bands for ROC curves", *Journal of Statistical Planning and Inference 138*, p. 1894-1904.

[6] **Lloyd, C. J.** (1998), "Using Smoothed Receiver Operating Characteristic Curves to Summarize and Compare Diagnostic Systems", *Journal of the American Statistical Association, Vol. 93, No. 444*, p. 1356-1364.

[7] **Lloyd, C. J., Zhou, Y.** (1999), "Kernel estimators of the ROC curve are better than empirical", *Statistics & Probability Letters 44*, p. 221-228.

[8] **R Development Core Team** (2010), *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

[9] **Serfling, R. J.** (1980), *Approximation Theorems of Mathematical Statistics*. John Wiley & Sons, Inc., New York.

[10] **Shorack, G. R. , Wellner, J. A.** (1986), *Empirical Processes With Applications to Statistics*. John Wiley & Sons, Inc., New York.

[11] **Zhou, X-H., Harezlak, J.** (2002), "Comparison of bandwidth selection methods for kernel smoothing of ROC curves", *Statistics in Medicine, Vol. 21*, p. 2045-2055.

[12] **Zou, K. H., Hall, W. J., Shapiro, D. E.** (1997), "Smooth Non-Parametric Receiver Operating Characteristic (ROC) Curves for Continuous Diagnostic Tests", *Statistics in Medicine, Vol. 16*, p. 2143-2156.

[13] **Zou, K. H., Hall, W. J.** (2000), "Two Transformation Models for Estimating an ROC curve Derived From Continuous Data", *Journal of Applied Statistics, Vol. 27, No. 5*, p. 621-631.

Apêndice A

Códigos para simulações em R

```
library(sn)
library(nor1mix)
library(KernSmooth)

    ##########################################
    #kernels for the distribution function
    ##########################################

    bnucleo.unif<-function(x)
    {punif(x,-1,1)}

    bnucleo.trian<-function(x)
    {(1/2)*((1+2*x+x^2)*(-1<x)*(x<=0) + (1+2*x-x^2)*(0<x)*(x<=1)) + 1*(x>1)}

    bnucleo.epane<-function(x)
    {(1/4)*(2 + 3*x - x^3)*(-1<x)*(x<1) + 1*(x>1)}

    bnucleo.biwei<-function(x)
    {(1/16)*(8 + 15*x - 10*x^3 + 3*x^5)*(-1<x)*(x<1) + 1*(x>1)}

    bnucleo.triwei<-function(x)
    {(1/32)*(16 + 35*x - 35*x^3 + 21*x^5 - 5*x^7)*(-1<x)*(x<1) + 1*(x>1)}

    ##################################################
    #Distribution function kernel estimador
    ##################################################

    # xx - pontos onde se pretende calcular o estimador
    # x - amostra
    # [a,b] - suporte de f; neste caso o estimador produzido é o estimador
    # com correcção de fronteira baseado nos núcleos de fronteira K(u/alpha)/alpha
    # Os casos a=-Inf e b=Inf correspondem ao estimador sem correcção de fronteira
    # Os casos a=-Inf ou b=Inf correspondem a correcções à direita ou à esquerda
```

```
    dfke <- function(xx,x,h,a=-Inf,b=Inf,kernel="epane")
    {
     if (kernel=="unif") bK <- bnucleo.unif else
     if (kernel=="trian") bK <- bnucleo.trian else
     if (kernel=="epane") bK <- bnucleo.epane else
     if (kernel=="biwei") bK <- bnucleo.biwei else
     if (kernel=="triwei") bK <- bnucleo.triwei

     #estimador com correcção de fronteira
     xxl <- xx[xx<=a]
     yyl <- rep(0,times=length(xxl))
     xxlf <- xx[(xx>a)&(xx<a+h)]
     difl <- t(outer(xxlf, x, "-")/(xxlf-a))
     yylf <- colMeans(bK(difl))
     xxi <- xx[(xx>=a+h)&(xx<=b-h)]
     difi <- t(outer(xxi, x, "-")/h)
     yyi <- colMeans(bK(difi))
     xxrf <- xx[(xx>b-h)&(xx<b)]
     difr <- t(outer(xxrf, x, "-")/(b-xxrf))
     yyrf <- colMeans(bK(difr))
     xxr <- xx[xx>=b]
     yyr <- rep(1,times=length(xxr))
     est <- list(x=c(xxl,xxlf,xxi,xxrf,xxr),y=c(yyl,yylf,yyi,yyrf,yyr))

     return(est)
    }

#######Inversa Nucleo#########################################
invF0n<-function(F0n,p){
s<-0
p1<-which(0==F0n$y, arr.ind=TRUE)
p2<-which(1==F0n$y, arr.ind=TRUE)
p3<-0
for(i in 1:length(p)){
if(p[i]==0) s[i]<-F0n$x[length(p1)] else
if(p[i]==1) s[i]<-F0n$x[p2[1]] else{
p3<-which(p[i]<F0n$y, arr.ind=TRUE)
s[i]<-F0n$x[p3[1]]
}
}
return(s)
}
```

```
    ######################################################################
    #Calcula o integral de inf a sup de yy utilizando a formula
    #de Simpson composta
    #ordenadas calculadas num conjunto de num (ímpar) pontos
    #igualmente espaçados sendo inf e sup o primeiro e último
    #desses pontos
    ######################################################################

    integravector<-function(yy,inf=0,sup=1) #length(yy)=impar
    {
    num<-length(yy)
    passo<-(sup-inf)/(num-1)
    yy1<-yy[1:(num-1)]
    oo1<-rep(c(0,1),times=(num-1)/2)
    s4<-sum(oo1*yy1)
    yy2<-yy[2:(num-2)]
    oo2<-rep(c(0,1),times=(num-1)/2-1)
    s2<-sum(oo2*yy2)
    return(passo*(yy[1]+4*s4+2*s2+yy[num])/3)
    }
    xx <- seq(0,1,by=0.01)
    yy <- sin(xx)
    integravector(sin(xx))

########### Funcao Distribuicao Empirica##############################
fde<-function(X,p){sum(X<=p)/length(X)}

#####################################################################
###Transformacao Box-Cox
#####################################################################
  BoxCoxT<-function(X0,X1){
m<-length(X1)
n<-length(X0)
logVer<-function(l){
 y<-(m/2)*log(1/(2*pi*sd(((X1^l)-1)/l)^2))+(n/2)*log(1/(2*pi*sd(((X0^l)-1)/l)^2))
   -((1/(2*sd(((X1^l)-1)/l)^2))*sum((((X1^l)-1)/l)-mean(((X1^l)-1)/l))^2)
   -((1/(2*sd(((X0^l)-1)/l)^2))*sum((((X0^l)-1)/l)-mean(((X0^l)-1)/l))^2)
   +((l-1)*sum(log(X1)))+((l-1)*sum(log(X0)))}
lambda<-optimize(logVer, interval = c(-10,10), maximum=TRUE)$maximum
return(lambda)}

###############AUC-Estimativa Mann-Whitney#####################
```

```
   AUCMW<-function(X0,X1){
s<-0
for(j in 1:length(X0)){for(i in 1:length(X1))
{s<-s+(X1[i]>X0[j])+(0.5*(X1[i]==X0[j]))}}
return(s/(length(X0)*length(X1)))}

#########################################################################
############               Rotina de estimação                   #####
############Neste caso encontra-se exemplificado caso da lei normal #####
############assimétrica. Para os restantes casos basta adaptar com  #####
############as leis de probabilidade desejadas                      #####
#########################################################################

xx<-seq(0,10,length=10001)
p<-seq(0,1,length=10001)
ROC<-function(p){1-(psn(qsn(1-p,4.5,sqrt(0.09),15),5,sqrt(0.25),15))}
ROCV<-ROC(p)
plot(p,ROCV,type="l",lwd=2,col="black")
AUC<-integrate(ROC,0,1)

Estimar2<-function(xx,p,AUC,ROCV){
  p<-seq(0,1,length=10001)
  A<-0
  B<-0
  C<-0
  D<-0
  Erro1<-0
  Erro2<-0
  Erro3<-0
  Erro4<-0
  for(i in 1:1000){
X0<-rsn(200,4.5,sqrt(0.09),15)
X1<-rsn(200,5,sqrt(0.25),15)

###Estimacao da curva usando modelo normal
m0<-mean(X0)
m1<-mean(X1)
  s0<-sd(X0)
  s1<-sd(X1)
  ROCN<-function(p){(pnorm((s0/s1)*qnorm(p)+((m1-m0)/s1)))}
ROCEN<-ROCN(p)
Erro1[i]<-integravector(((ROCEN-ROCV)^2)
```

```
*dsn(qsn(1-p,4.5,sqrt(0.09),15),4.5,sqrt(0.09),15))
AUCN<-pnorm((m1-m0)/sqrt(s1^2+s0^2))
AUCMW1<-AUCMW(X0,X1)
A[i]<-(AUC$value - AUCN)
C[i]<-(AUC$value - AUCMW1)

###Normalizacao dos dados e estimacao da curva usando modelo normal
lambda<-BoxCoxT(X0,X1)
B0<-(X0^(lambda)-1)/lambda
B1<-(X1^(lambda)-1)/lambda
muu0<-mean(B0)
muu1<-mean(B1)
sigmaa0<-sd(B0)
sigmaa1<-sd(B1)
ROCB<-function(p){(pnorm((sigmaa0/sigmaa1)*qnorm(p)
+((muu1-muu0)/sigmaa1)))}
ROCEB<-ROCB(p)
Erro2[i]<-integravector(((ROCEB-ROCV)^2)
*dsn(qsn(1-p,4.5,sqrt(0.09),15),4.5,sqrt(0.09),15))
AUCNB1<-pnorm((muu1-muu0)/sqrt(sigmaa1^2+sigmaa0^2))
B[i]<-(AUC$value - AUCNB1)

###Estimacao nao parametrica da curva
ROCNP<-0
for(j in 1:length(p)){
ROCNP[j]<- 1-(fde(X1,quantile(X0,1-p[j])))}
Erro3[i]<-integravector(((ROCNP-ROCV)^2)
*dsn(qsn(1-p,4.5,sqrt(0.09),15),4.5,sqrt(0.09),15))

###Estimacao do nucleo
a<-((4*sqrt(pi)*(0.2571429))/(1/25))
b0<-((s0^3)*(((s0^2)*(s1^2))^(5/2)))/((((s1^2)+(2*(s0^2)))^(1/2))
*((s1^4)+(s1^2*s0^2)+(2*(s0^2)*((m0-m1)^2))))
c0<-exp(((m0-m1)^2*(s0^2))/((s0^2+s1^2)*(2*s0^2+s1^2)))
b1<-((s1^3)*(((s1^2)*(s0^2))^(5/2)))/((((s0^2)+(2*(s1^2)))^(1/2))
*((s0^4)+(s0^2*s1^2)+(2*(s1^2)*((m1-m0)^2))))
c1<-exp(((m1-m0)^2*(s1^2))/((s1^2+s0^2)*(2*s1^2+s0^2)))
d0<-(a*b0*c0)^(1/3)
h0<-d0*(length(X0)^(-1/3))
F0n<-dfke(xx,X0,h0,a=-Inf,b=Inf,kernel="epane")
d1<-(a*b1*c1)^(1/3)
h1<-d1*(length(X1)^(-1/3))
```

```
F1n<-dfke(xx,X1,h1,a=-Inf,b=Inf,kernel="epane")
F00<-invF0n(F0n,1-p)
F11<-dfke(F00,X1,h1,a=-Inf,b=Inf,kernel="epane")
ROCk<-1-F11$y
AUCk<-integravector(ROCk)
Erro4[i]<-integravector(((ROCk-ROCV)^2)
*dsn(qsn(1-p,4.5,sqrt(0.09),15),4.5,sqrt(0.09),15))
D[i]<-(AUC$value - AUCk)
}
png("AUC.png")
boxplot(A,B,C,D)
dev.off()

png("Erros.png")
boxplot(Erro1,Erro2,Erro3,Erro4)
dev.off()

medidasAUC<-c(mean(A),sd(A),mean(B),sd(B),mean(C),sd(C),mean(D),sd(D))
medidasErro<-c(mean(Erro1),sd(Erro1),mean(Erro2),sd(Erro2),mean(Erro3)
,sd(Erro3),mean(Erro4),sd(Erro4))

tabela1<-matrix(medidasAUC,nrow=4,ncol=2,byrow=TRUE)
tabela2<-matrix(medidasErro,nrow=4,ncol=2,byrow=TRUE)

write.table(tabela1, file="AUCNorA.txt")
write.table(tabela2, file="ErrosNorA.txt")
}
Estimar(xx,p,AUC,ROCV)
```