Universidade Federal de Minas Gerais
Instituto de Ciências Exatas
Departamento de Estatística
Especialização em Estatística

Nucleo-estimador da função de densidade e da função de distribuição.

Wagner Luiz Moreira dos Santos

Belo Horizonte-MG

Novembro de 2014

Wagner Luiz Moreira dos Santos

Nucleo-estimador da função de densidade e da função de distribuição.

Monografia do curso de especialização em estatítica apresentada ao Departamento de Estatística do Instituto de Ciências Exatas da Universidade Federal de Minas Gerais como requisito parcial para a obtenção do grau de especialista em Estatística.

Orientador:

Gregorio Saravia Atuncar

Universidade Federal Minas Gerais – UFMG
Departamento de Estatística

Belo Horizonte-MG

Novembro de 2014

Agradecimentos

Gostaria de iniciar esta seção agradecendo ao professor Gregorio pela dedicação, orientação e infinita paciência dispensados no decorrer deste trabalho; ao professor Ricardo Tavares (UFOP) pela disponibilidade, apoio e auxílio sempre quando solicitado; à minha amiga, esposa e companheira Ana Tereza Lana que foi meu porto seguro em vários momentos difícies da minha vida e principalmente à minha mãe por seu exemplo e amor.

Nucleo-estimador da função de densidade e da função de distribuição.

Autor: Wagner Luiz Moreira dos Santos
Orientador: Prof. Gregorio Saravia Atuncar

Resumo

Neste trabalho discutimos algumas técnicas para a estimação da função de densidade e da função de distribuição utilizando Núcleo-estimador, mostrando suas principais vantagens e desvantagens. Analisamos as propostas para o cálculo da "*optimal bandwidth*" ou como definido em português, "janela ótima". Numa primeira fase, visando analisar o desempenho do método, simulamos amostras de variáveis aleatórias com várias funções de densidade para diferentes valores de parâmetros. Estimamos a função de densidade/distribuição utilizando o Núcleo-estimador e comparamos com as funções de densidade/distribuição reais correspondentes. Quando a variância amostral das observações $x_1, x_2, \cdots, x_n$, é extremamente grande ou muito pequena este método não é apropriado, por este motivo, realizamos uma transformação linear "Y=aX", de tal forma que a variância amostral dos y_i´s seja 1. Assim estimamos a função de densidade/distribuição de "Y" e posteriormente a função de densidade/distribuição de "X". Finalmente abordamos um exemplo de aplicação no qual tentamos ajustar várias funções paramétricas sendo então necessária a aplicação do Núcleo-estimador para estimar a densidade geradora dos dados amostrais.

Palavras-chave: *Núcleo-estimador, janela ótima, transformação linear.*

Lista de figuras

Lista de tabelas

Sumário

1 Introdução

Uma idéia básica em Estatística é o conceito de função de densidade de probabilidade. Esta função tem como objetivo modelar estatísticamente a probabilidade através dos parâmetros da distribuição. Apesar da grande importância desta função, em algumas situações podemos ter como objetivo principal a forma da função de densidade ou distribuição.

O histograma é o estimador da função densidade mais utilizados. Estas ferramentas mostram o comportamento distributivo de um conjunto de dados. No histograma, a área da barra é proporcional ao número de observações no intervalo ao qual pertence. A escolha do comprimento (amplitude) desses intervalos controla a suavidade do estimador.

Segundo (LUCAMBIO, 2008), a partir de 1890, diferentes formas de estimar uma função de densidade tem sido propostas. Uma destas é devida a Karl Pearson (1857-1900) sendo obtida como solução de uma equação diferencial, veja (JOHNSON; KOTZ, 1988).

A partir de 1956 os métodos não-paramétricos de estimação de funções de densidade de probabilidade têm se consolidado como uma alternativa sofisticada para o estudo de conjuntos de dados, podendo assim, analisar esses conjuntos sem conhecimento da regra de correspondência para a função de densidade ou função de distribuição.

Este tema tem despertado muito interesse em pesquisadores de todas as áreas do conhecimento, devido à sua ampla abordagem. Até a presente data já foram publicados milhares de artigos sobre o tema e espera-se ainda a publicação de muitos outros, já que ainda existem muitas questões controvérsas ou ainda não resolvidas sobre o tema.

Núcleo estimador é uma técnica não-paramétrica para estimação de funções, fornecendo o estimador da função de densidade de probabilidade, ou distribuição, de uma forma atraente e sem a imposição de um modelo paramétrico. O desempenho do núcleo-estimador depende essencialmente da escolha do parâmetro de suavidade. Este parâmetro, que chamaremos de janela e na literatura em inglês é denominada "*bandwidth*", geralmente denotado por "h_n", determina o grau de suavização a ser feita, isto é, ela é responsável por

quão rapidamente o estimador da função oscila. Para estimarmos, de maneira ótima, as funções de densidade "f" ou distribuição "F", devemos encontrar o tamanho ideal deste parâmetro ($janela$). Neste trabalho, a janela "h_n" será estimada através do método *Plug-in* modificado que será descrito posteriormente com mais detalhes.

Nesta monografia, mostraremos apenas o caso da janela global, ($fixa$), apresentando alguns métodos para estimar o melhor valor desta janela baseado na definição do núcleo estimador. O caso da janela variável não será abordado aqui, podendo ser analisado em trabalhos futuros. Na Figura 1, podemos verificar a necessidade, ou não, do uso da da janela variável. Verifique que no gráfico de linha sólida o número de elementos dentro de cada intervalo varia muito, e rapidamente, no intervalo $(0, 10)$, justificando o uso da janela variável; o que não ocorre, tão rapidamente no gráfico de linhas pontilhadas.

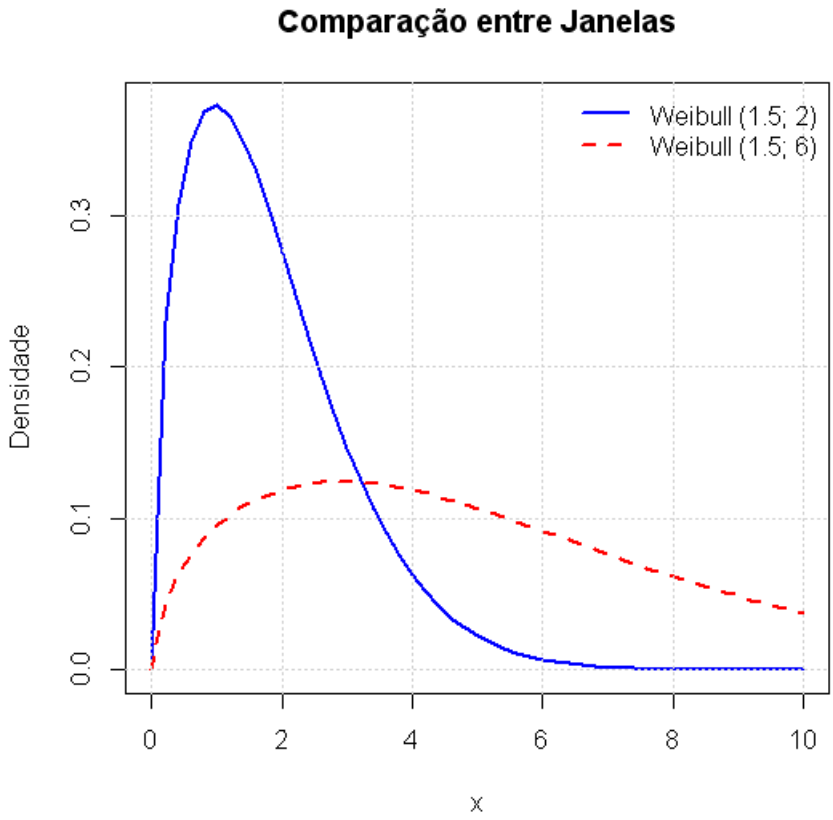

Figura 1: No caso (—) é necessário o uso de janela variável já que o número de elementos dentro dos intervalos de comprimentos iguais varia muito. E no caso (- - -) usamos a janela global já que há pouca variação no número de elementos nos diferentes intervalos.

1.1 Revisão Bibliográfica

(FIX; HODGES, 1951) foram os primeiros autores na literatura que propuseram as ideias básicas sobre Núcleo-estimador, utilizando uma função núcleo Uniforme no intervalo (0,1). (ROSENBLAT, 1956) e (PARZEN, 1962) estudaram a classe geral do núcleo estimador univariado. (ROSENBLAT, 1956), mostrou que todo estimador é viciado para a função de densidade de probabilidade, o que motivou a procura de estimadores assintoticamente não viciados. (PARZEN, 1962), examinou as propriedades assintóticas do núcleo estimador em termos do erro quadrático médio (EQM) e mostrou que, sobre algumas condições, o núcleo estimador é assintoticamente não-viciado. Outros resultados sobre propriedades do núcleo estimador podem ser encontrados em (BERTRAND-RETALI, 1978) e (DEVROYE; GYORFI, 1978).

(EPANNECHNIKOV, 1969) fez um estudo da eficiência deste estimador, em termos assintóticos, de diferentes funções núcleo, tendo como objetivo encontrar uma função núcleo ótima. O mesmo problema foi considerado por (BARTLETT, 1963), entretanto, (SILVERMAN, 1986) e (TERRELL; SCOTT, 1985) mostraram que o desempenho do núcleo estimador é dominado essencialmente pela escolha do parâmetro de suavidade.

Um dos primeiros métodos automáticos [1] de estimação da janela ótima foi o método *"plug-in"* proposto por (WOODROOFE, 1970) e (NADARAYA, 1974). (HABBEMA; HERMANS; BROEK, 1974) e (DUIN, 1976) propuseram, independentemente, o método de validação cruzada por verossimilhança para a escolha da janela ótima. Entretanto, este método pode produzir estimativas inconsistentes, como foi observado no trabalho de (SCHUSTER; GREGORY, 1981).

(RUDEMO, 1982) e (BOWMAN, 1984), desenvolveram, independentemente, o método de validação cruzada por mínimos quadrados (VCMQ). No entanto, este método possui alta variabilidade, o que compromete seu desempenho. Na tentativa de melhorar este método, (TERRELL; SCOTT, 1985) propuseram o método de validação cruzada viciada (VCV), que mostrou-se mais estável que o (VCMQ) no que se refere à variação do método. Entretanto, esta redução da variância implica em um aumento do vício, o que foi comparado por (JONES; KAPPENMAN, 1992) em ambas as metodologias.

Nesta monografia utilizaremos o método *plug-in* modificado, proposto por Chiu, para estimar a janela ótima. Para uma descrição mais aprofundada sobre o tema, veja (SILVERMAN, 1986), (TERRELL; SCOTT, 1985), (WAND; JONES, 1994), (NADA,) ou (BOWMAN;

[1] métodos baseados na idéia de que o parâmetro de suavização deve depender unicamente dos dados

AZZALINI, 1997).

1.2 Objetivos e Organização do Trabalho

Nesta monografia temos como objetivo principal a análise da técnica de Núcleo-estimador, simulação e a aplicação a uma base de dados reais. Na parte inicial fizemos um breve análise sobre o tema, dentre os principaís autores da área. Já na parte de simulação analisamos a eficácia do estimador comparando a função real com sua função estimada e analisando o resíduo gerado pelo mesmo.

No capítulo 2 fizemos uma análise da função de densidade. Iniciamos definindo "janela ótima" e comparamos gráficos para diferentes tamanhos de janelas. Mostramos algumas técnicas para a estimação deste parâmetro. Analisamos o comportamento da janela ótima para vários desvios populacionais simulados. Verificamos também o comportamento da janela ótima quando variamos seu tamanho amostral "N". Finalmente fazemos uma comparação gráfica de amostras com distribuições normal e Weibull via Núcleo-estimador com sua função de densidade real.

No capítulo 3 realizamos as comparações, feitas no capítulo 2 para a função de densidade, para a função de distribuição acumulada.

No capítulo 4 mostramos que quando temos variância muito grande ou muito pequena temos a necessidade em realizar uma transformação linear devido ao decaimento muito rápido ou muito lento da ($F.C.E$). Ainda no capítulo 4, mas na parte de aplicação, aplicamos a técnica a uma base de dados, tentando ajustar sua densidade à funções paramétricas conhecidas e posteriormente ajustando via Núcleo-estimador.

2 Núcleo-estimador da Função de Densidade.

O Núcleo estimador é uma técnica que tem como objetivo principal a estimação de uma função de distribuição ou densidade, diferentemente dos modelos paramétricos. O Núcleo-estimador fornece um caminho simples para encontrar a estrutura probabilística de um conjunto de dados. Nesta monografia iremos trabalhar apenas com funções de densidae e fuções de disribuição univariadas, contínuas e o Núcleo-estimador será usado apenas em janela fixa.

Esta técnica depende da escolha de um parâmetro de suavidade h_n do estimador, sendo chamado na literatura em inglês como *bandwidth*, que neste texto chamaremos de "janela". Assim, a estimação desta janela é uma parte fundamental no processo de estimação das funções de densidade ou distribuição (SILVERMAN, 1986).

Considere uma amostra aleatória de tamanho n definida por X_1, X_2, ..., X_n de uma variável aleatória X com função de densidade f. O núcleo estimador de f, definido por (ROSENBLAT, 1956), no ponto x é:

$$f_n(x) = \frac{1}{nh_n} \sum_{i=1}^{n} k\left(\frac{x_j - X_i}{h_n}\right) \tag{2.1}$$

onde "k" é uma função de densidade, sendo chamada de núcleo.

Pode-se provar que se $h_n \to 0$ e $n.h_n \to \infty$, então, $f_n(x) \xrightarrow{p} f(x)$. Mas existe uma infinidade de sequências h_n com essa propriedade. Para valores pequenos de h_n dizemos que o parâmetro de suavidade é "aspero" e para valores grandes de h_n dizemos que este parâmetro é "muito macio". Uma exemplificação é apresentada na ($Figura2$).

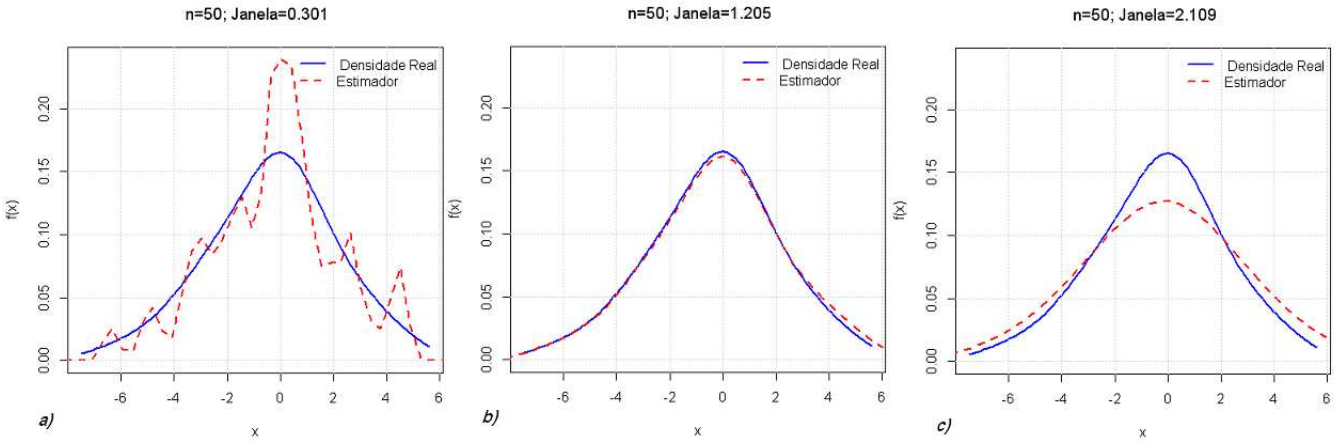

Figura 2: Densidade estimada de uma normal (0, 4) de tamanho n = 50. A linha sólida indica a densidade real e a linha tracejada indica a densidade estimada via núcleo estimador. (a) $h_n = 0.301$, (parâmetro áspero), (b) $h_n = 1,205$, (parâmetro ótimo) e (c) $h_n = 2.109$, (parâmetro muito suave)

2.1 Estimação da janela ótima

A estimação do tamanho da janela ótima ou parâmetro de suavidade "h_n", é o passo mais importante na estimação das funções de densidade via núcleo estimador, devido ao fato desta escolha ser diretamente responsável pelo controle de suavidade de $\hat{f}_n(x)$.

Existem algumas técnicas para o cálculo deste estimador, dentre elas podemos citar o metodo de validação cruzada, proposta separadamente por (RUDEMO, 1982) e (BOWMAN, 1984) além do *Plug-in*, proposto inicialmente por (WOODROOFE, 1970). Hall e Marron (1987) observaram que o método de validação cruzada e o método *Plug-in*, fornecem estimadores com grande variabilidade. Assim (CHIU, 1991) propos o método *Plug-in* modificado que controla esta variabilidade.

A partir da definição que será utilizada no Erro Quadrático Médio Integrado (EQM), temos que a janela ótima é dada por:

$$h_{opt} = \left[\frac{\int k^2(z)\,dz}{n\left[\int z^2 k(z)\,dz\right]^2 \int |f''(x)|^2 dx} \right]^{\frac{1}{5}} \tag{2.2}$$

Um problema que aparece nesta fórmula, é que a janela ótima depende diretamente do termo $\int |f''(x)|^2 dx$, sendo $f''(x)$ a derivada segunda da função de densidade de probabiliade que é uma função desconhecida para nós. Para mais detalhes veja (ATUNCAR; DAMASCENO; MENDONçA, 2008) e (SILVERMAN, 1986).

Alguns autores propuseram técnicas para a solução deste problema, dentre eles podemos citar (WOODROOFE, 1970) que propôs o método Plug-in. Este método utiliza uma aproximação para o único termo desconhecido na expressão, ou seja a $\int |f''(x)|^2 dx$. O autor utiliza um estimador prévio de "f' usando um "h_n" apropriado e, substituindo na integral, pela aproximação da janela ótima. O problema é que encontrar este estimador prévio não é uma tarefa muito fácil.

(CHIU, 1991), através do método *Plug-in* modificado, substituiu este termo por uma aproximação de "G", onde $G = \int |f''(x)|^2 dx$. Utilizando a função característica, temos:

$$f''(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{i\lambda x} \left(-\lambda^2\right) |\varphi(\lambda)|^2 d\lambda$$

Usando a identidade de Parseval, veja (CHIU, 1991), podemos provar que:

$$G = \int_{-\infty}^{\infty} [f''(x)]^2 \, dx = \frac{1}{2\pi} \int_{-\infty}^{\infty} \lambda^4 |\varphi(\lambda)|^2 d\lambda$$

Assim, podemos aproximar G por:

$$\hat{G} = \frac{1}{\pi} \int_0^{\Lambda} \lambda^4 \left[|\hat{\varphi}(\lambda)|^2 - \frac{1}{n} \right] d\lambda \tag{2.3}$$

onde $\Lambda = min\left\{\lambda : |\hat{\varphi}(x)|^2 \leq \frac{c}{n}\right\}$

Chiu mostrou que a função estimada $\hat{\varphi}_x(\lambda)$ domina seu erro no intervalo $[0, \Lambda]$. (DAMASCENO, 2000) verificou através de simulações de grande porte, que $c = 3$ reduz a variabilidade do estimador. Devido à este resultado, escolhemos $c = 3$ para a função estimada em (2.3). Para maiores detalhes veja (ATUNCAR; DAMASCENO; MENDONçA, 2008) e (BESSEGATO, 2001).

2.2 Método de Validação Cruzada

Suponha um estimador $\hat{f}$ da função de densidade de probabilidade f de uma variável aleatória. Por definição, o *"erro quadrático integrado"* EQI é dado por:

$$EQI_n(f_n) = \int \left[\hat{f}_n(x) - f(x)\right]^2 dx = \int \hat{f}_n^2(x)dx - 2\int \hat{f}_n(x)f(x)dx + \int f^2(x)dx \tag{2.4}$$

Podemos observar que o último termo não depende do estimador de $\hat{f}_n$. Assim, minimizar o EQI_n, é equivalente a minimizar a expressão acima sem o último termo, ou seja:

$$CV = \int \hat{f}_n^2(x)dx - 2\int \hat{f}_n(x).f(x)dx \tag{2.5}$$

No entanto podemos perceber que a equação (2.5) ainda depende de "f". Como esta função é desconhecida, podemos estima-la aravés dos próprios dados.

O método de validação cruzada, para a estimação da janela ótima "h'', foi proposto separadamente por (RUDEMO, 1982) e (BOWMAN, 1984). Segundo estes autores, este estimador é dado por:

$$CV_n(h_n) = \int \hat{f}^2(x)dx - \frac{2}{n}\sum_{i=1}^{n} f_i(X_i) \tag{2.6}$$

sendo,

$$f_i(x_i) = \frac{1}{(n-1)h_n}\sum_{j\neq i} K(\frac{X_i - X_j}{h_n}) \tag{2.7}$$

Assim, podemos verificar que a equação (2.6) depende apenas dos dados amostrais.

O método de validação cruzada foi o pioneiro para a estimação da janela ótima. Neste trabalho utilizaremos o método *Plug-in* modificado, que devido ao seu balanceamento entre a variância e vício gera melhores resultados comparado com o método de validação cruzada e o método *Plug-in*.

2.3 O Método Plug-in Modificado

Este método, diferentemente do *Plug-in* original proposto por (WOODROOFE, 1970) não necessita de uma escolha inicial de h para posteriormente chegar à um $(h - otimo)$

Um problema existente para o cálculo da janela ótima, é que sua fórmula depende da derivada segunda da função de densidade f, que ainda será estimada. Veja equação (2.2). Sendo assim, Chiu propôs o método *Plug-in* modificado que estima a quantidade desconhecida $\left(\int |f''(x)|^2 dx\right)$ na fórmula de h_{opt} e define como janela ótima o argumento na minimização do Erro Quadrático Médio Integrado.

$$\begin{aligned} EQMI &= \int E\left(f_n(x) - f(x)\right)^2 dx \\ &= \frac{1}{nh}\int k(t)^2 dt + O\left(\frac{1}{n}\right) + \frac{1}{4}h^4 k_2^2 \int f''(x)^2 dx + o\left(h^4\right) \end{aligned} \tag{2.8}$$

O erro quadrático médio é uma medida de acurácia do núcleo estimador. Assim, o erro quadrático médio avaliado em "x" é definido por:

$$\begin{aligned} EQM\left[\hat{f}(x)\right] &= E\left\{\left[f(x)-\hat{f}(x)\right]^2\right\} \\ &= E\left\{\hat{f}(x)-E\left[\hat{f}(x)\right]\right\}+\left\{E\left[\hat{f}(x)\right]-\hat{f}(x)\right\}^2 \\ &= Var\left[(x)\right]+\left\{vicio\left[\hat{f}(x)\right]\right\}^2 \end{aligned} \tag{2.9}$$

Diferenciando o Erro Quadrático Médio Integrado, em relação à "h" e igualando à zero, temos:

$$\frac{\partial(EQMI)}{\partial h}=0\Rightarrow h_{opt} \tag{2.10}$$

onde h_{opt} é dado por:

$$h_{opt}=\left[\frac{\int k^2(z)\,dz}{n\left[\int z^2k(z)\,dz\right]^2\int|f''(x)|^2dx}\right]^{\frac{1}{5}}$$

Que é a equação dada em (2.2).

2.4 As Funções Características.

As definições de função característica e função característica empírica encontram-se a seguir:

A função caracaterística $(F.C)$ de uma variável aleatória X, avaliada em λ, é definida como:

$$\varphi(\lambda)=\varphi_x(\lambda)=E\left[e^{i\lambda X}\right] \tag{2.11}$$

onde define-se:

$$E\left[e^{i\lambda X}\right]=E\left[cos(\lambda.X)\right]+i.E\left[sen(\lambda.X)\right]$$

com $\lambda\in\Re$.

Seja $X_1,\cdots,X_n$, uma amostra aleatória com função de distribuição "F" e seja φ a função característica de "F". Define-se como função característica empírica $(F.C.E)$, representada por $\hat{\varphi}$, avaliada em λ por:

$$\hat{\varphi}_x(\lambda) = \frac{1}{n}\sum_{j=1}^{n} e^{i\lambda X_j} \tag{2.12}$$

onde define-se:

$$E\left[e^{i\lambda X_j}\right] = E\left[cos\left(\lambda X_j\right)\right] + iE\left[sen\left(\lambda X_j\right)\right] \tag{2.13}$$

Aplicando-se propriedades de números complexos, temos:

$$\begin{aligned}\hat{\varphi}_x(\lambda) &= \frac{1}{n}\left[\sum_{j=1}^{n} cos\left(\lambda X_j\right) + i\sum_{j=1}^{n} sen\left(\lambda X_j\right)\right] \\ &= \frac{1}{n}\sum_{j=1}^{n} cos\left(\lambda X_j\right) + \frac{i}{n}\sum_{j=1}^{n} sen\left(\lambda X_j\right)\end{aligned} \tag{2.14}$$

Estas definições serão utilizadas nas seções seguintes.

2.5 Estimativa Plug-in proposta por Chiu:

Como já citado anteriormente, a única quantidade não conhecida na fórmula da janela ótima h_{opt} é a integral, definida por $G = \int \left(f''(x)\right)^2 dx$. Assim Chiu propõe a substituição da integral por uma estimativa de G dada por:

$$\hat{G} = \frac{1}{\pi}\int_0^{\Lambda} \lambda^4 \left[|\hat{\varphi}(\lambda)|^2 - \frac{1}{n}\right] d\lambda \tag{2.15}$$

onde $\Lambda = min\left\{\lambda : |\hat{\varphi}(x)|^2 \leq \frac{c}{n}\right\}$.

Entretanto, quando temos uma variável aleatória com desvio padrão σ muito pequeno a função característica "F.C" e a função característica empírica "F.C.E" decrescem muito lentamente, dificultando o encontrar o valor de Λ. A *Figura 3* mostra este decaímento muito lento, quando σ é muito pequeno, para a função característica da normal $(0, \sigma)$.

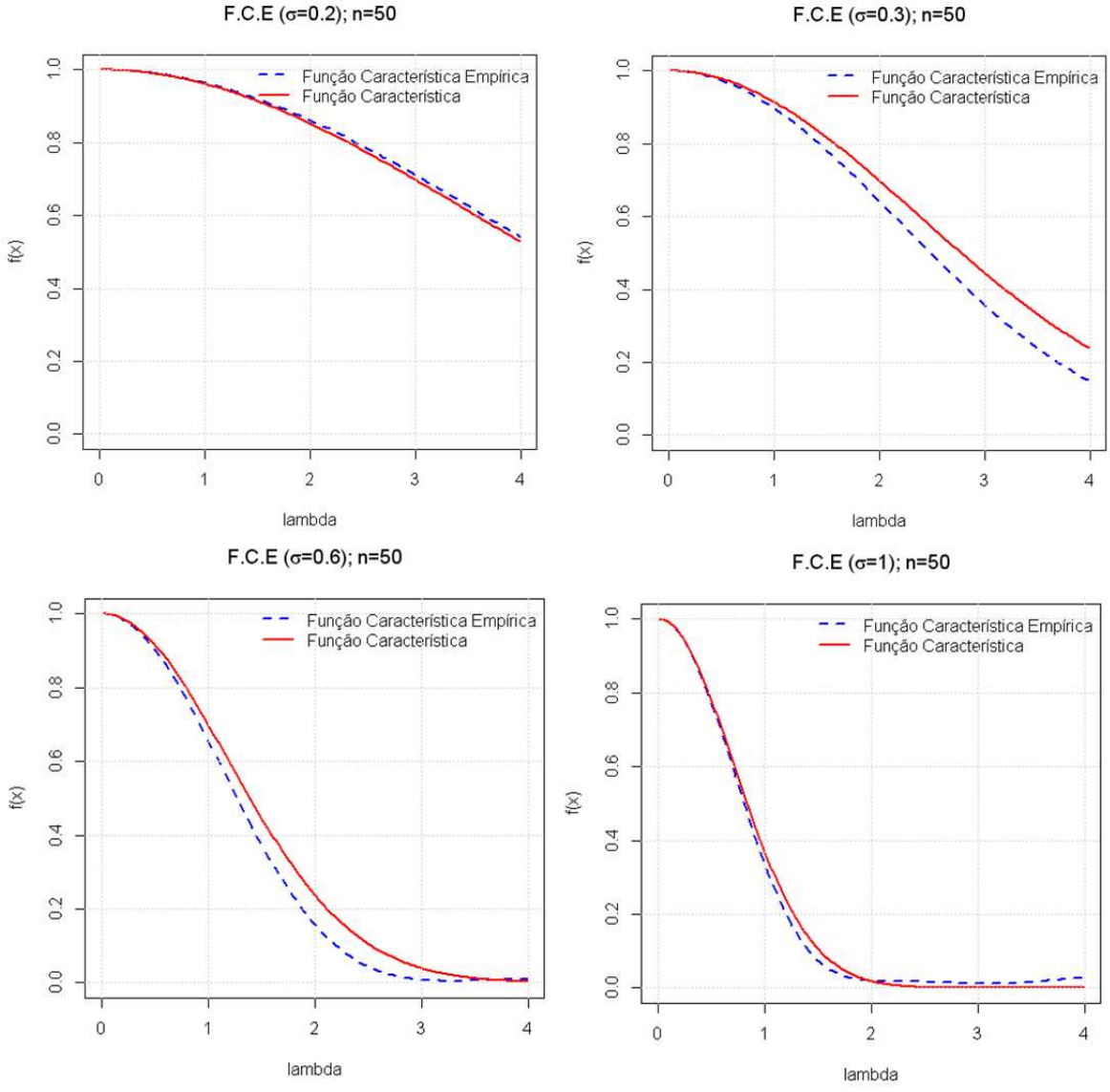

Figura 3: Análise da convergência "lenta" da função característica pelo desvio padrão σ - "pequeno".

Se por outro lado o desvio padrão σ é extremamente grande a função característica decresce muito rápido. Neste caso o cálculo da integral dada em (2.15) fica comprometida. Podemos verificar esta convergência rápida para a "F.C" e "F.C.E" na *Figura 4*.

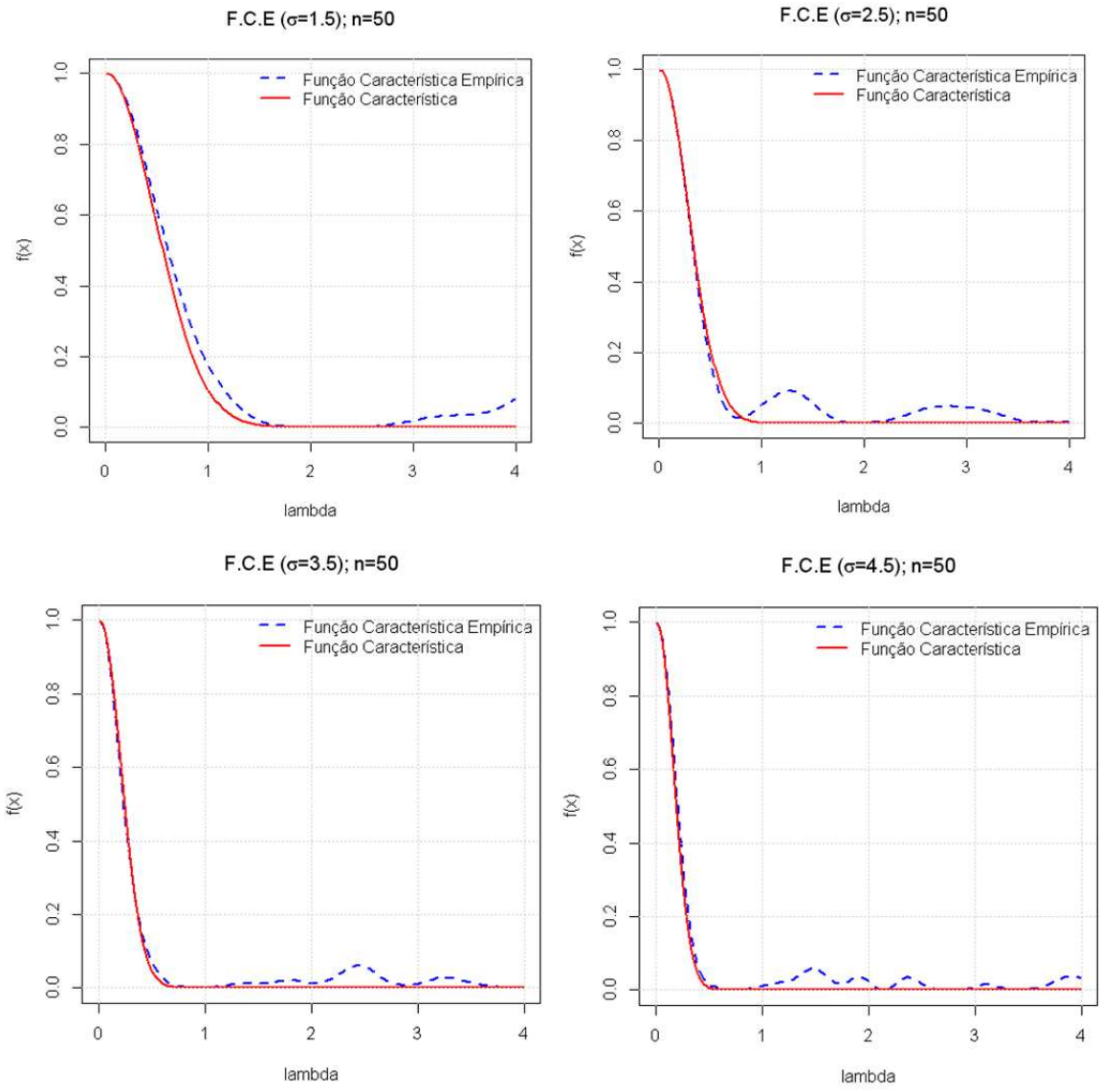

Figura 4: Análise da convergência "rápida" da função característica pelo desvio padrão σ - "grande".

Os gráficos das *Figuras 3 e 4*, mostram o decaímento lento ou rápido das funções características. Nestes casos, para conseguirmoso valor de Λ é recomendável uma utilizarmos uma transformação linear. Esta transformação divide todos os elementos da amostra por seu desvio amostral "s", determina seu núcleo estimador para posteriormente voltar à variável original. Mais detalhes desta transformação serão vistos no capítulo 4 deste trabalho.

(CHIU, 1991) mostra que a escolha do parâmetro "c" não é relevante quando "f" é suficientemente suave. (DAMASCENO, 2000), otimizou este parâmetro "c" através de simulações de grande porte, concluindo que o valor que melhor pondera o vício e a variabilidade é sempre igual a 3.

Assim, quando substituímos a integral $G = \int \left(f''(x)\right)^2 dx$ por $\hat{G}$ em (2.3) na fórmula da janela ótima, h_{opt}, temos que o estimador da janela ótima é dado por:

$$\hat{h}_{chiu} = \left[\frac{\int K^2(z)\, dz}{n \left[\int z^2 K(z)\, dz\right]^2 \hat{G}} \right]^{\frac{1}{5}} \tag{2.16}$$

que não necessita do conhecimento de "f".

2.6 Variação da Janela Ótima

Nesta seção observamos a relação entre o comprimento da Janela Ótima em relação à variação de seu desvio padrão e também a variação desta janela em relação ao número de elementos amostrais "n", utilizando a fórmula (2.16).

Análise 1 - Consiste em gerar 100 amostras de tamanho n=50 da distribuição normal com média $\mu = 0$, para cada valor de desvio padrão $\sigma = \{0,1; 0,2; 0,3; \cdots, 4,8; 4,9; 5,0\}$. Com estes valores, calculamos a média e o desvio padrão da janela ótima para cada um destes σ, fazendo o gráfico do desvio padrão -σ *versus* Janela Ótima.

Através do gráfico da *Figura 5*, podemos verificar que o desvio padrão σ e a Janela Ótima são diretamente proporcionais.

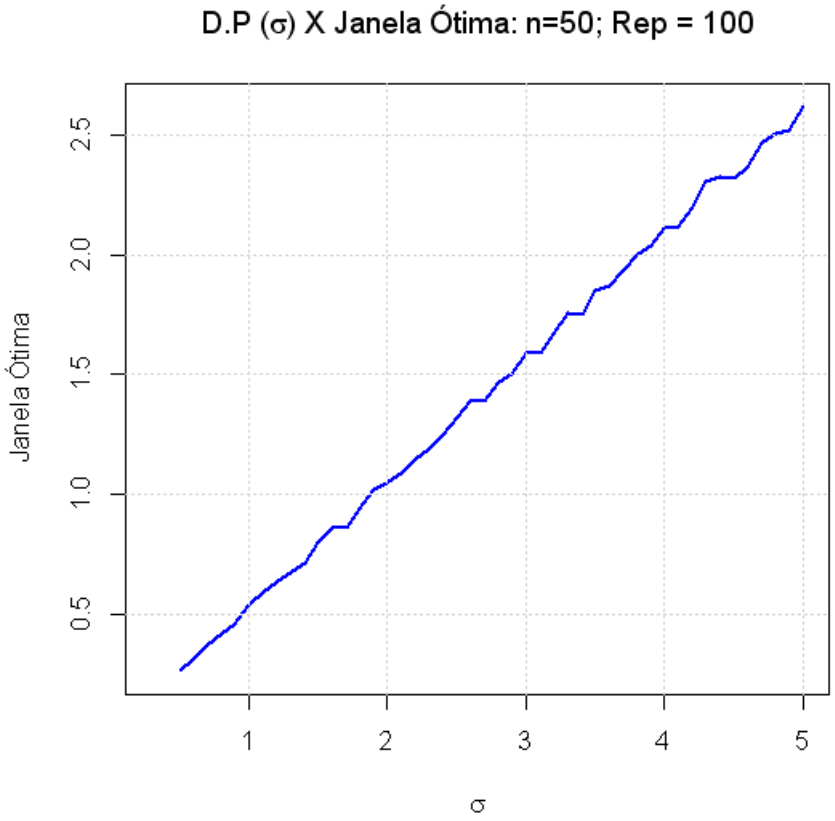

Figura 5: (Análise 1) - Variação da Janela Ótima em função do desvio padrão σ.

Análise 2 - Consiste em gerar 100 amostras normais de $\mu = 0$, $\sigma = 2$ com amostras de tamanho $n = \{30; 60; 90; \cdots ; 960; 990; 1020\}$. Através destes tamanhos amostrais verificamos a variação do comprimento da Janela Ótima em função do tamanho amostral "n", fazendo o gráfico -"n" *versus* Janela Ótima-

Podemos também verificar que a Janela Ótima é inversamente proporcional ao tamanho amostral "n", como mostra a *Figura 6*, ou seja quando $n \to \infty$ temos $h_n \to 0$.

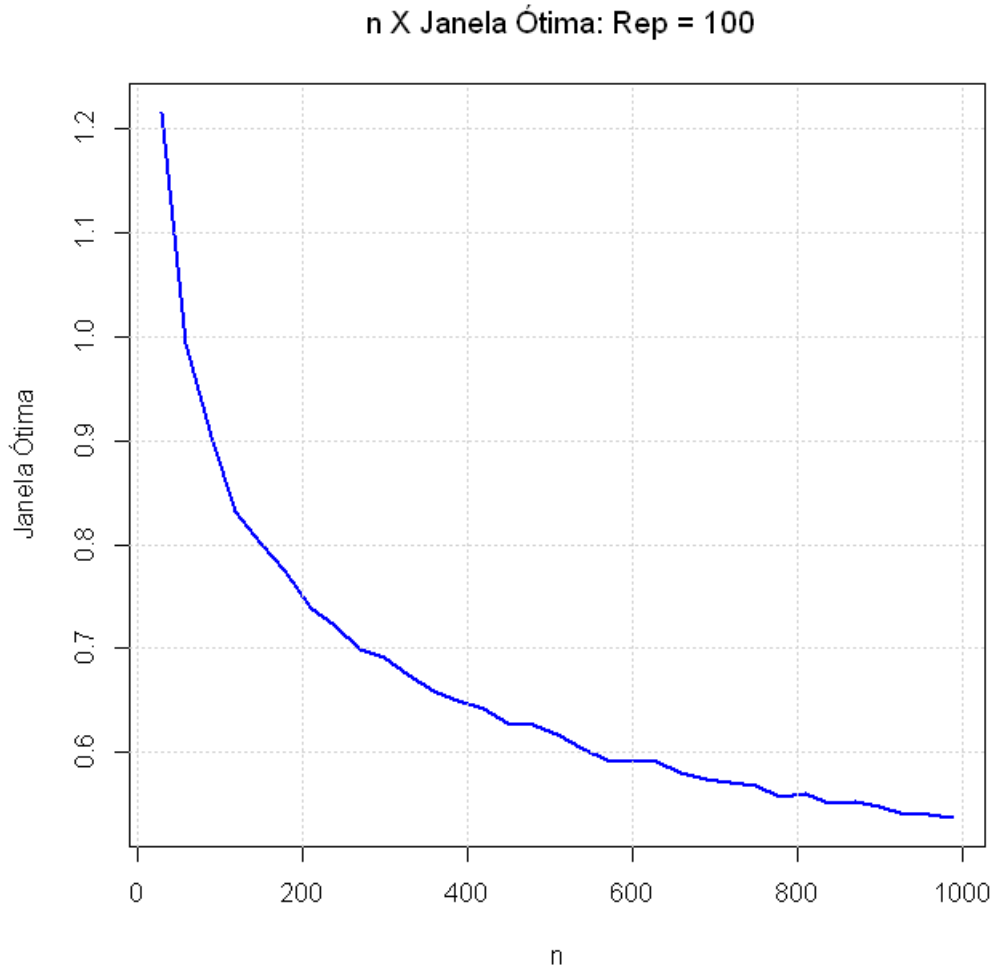

Figura 6: (Análise 2) - Variação da Janela Ótima em função do tamanho amostral "n".

2.7 Comparação Gráfica - Função Densidade

Nesta seção compararemos funções de densidade reais com as estimadas usando o núcleo estimador. Utilizaremos as distribuições Normal e Weibull para esta comparação, analisando assim o resíduo do estimador. Para ambos os casos verificamos que quanto maior é o valor de "n", menor será o resíduo do estimador, ou seja, melhor será a aproximação do estimador comparado com a função real, veja as *figuras 7 e 8*.

2.7.1 Distribuição Normal (0,4)

No *software "R"* geramos amostras com distribuição normal com média zero e desvio padrão igual a 2 para $n = \{50; 100; 500; 1000\}$. Estimamos sua densidade através das equações (2.1) e (2.2) e comparamos estas estimativas com a função real.

Podemos verificar, através da *Figura 7*, que o Núcleo-estimador tem menor resíduo quando o tamanho amostral "n" aumenta. Veja que para "n=1000" os dois gráficos são muito próximos. Mas também podemos observar que, para situações reais, quando "n=50" ou "n=100" a função estimada aproxima-se bem da real.

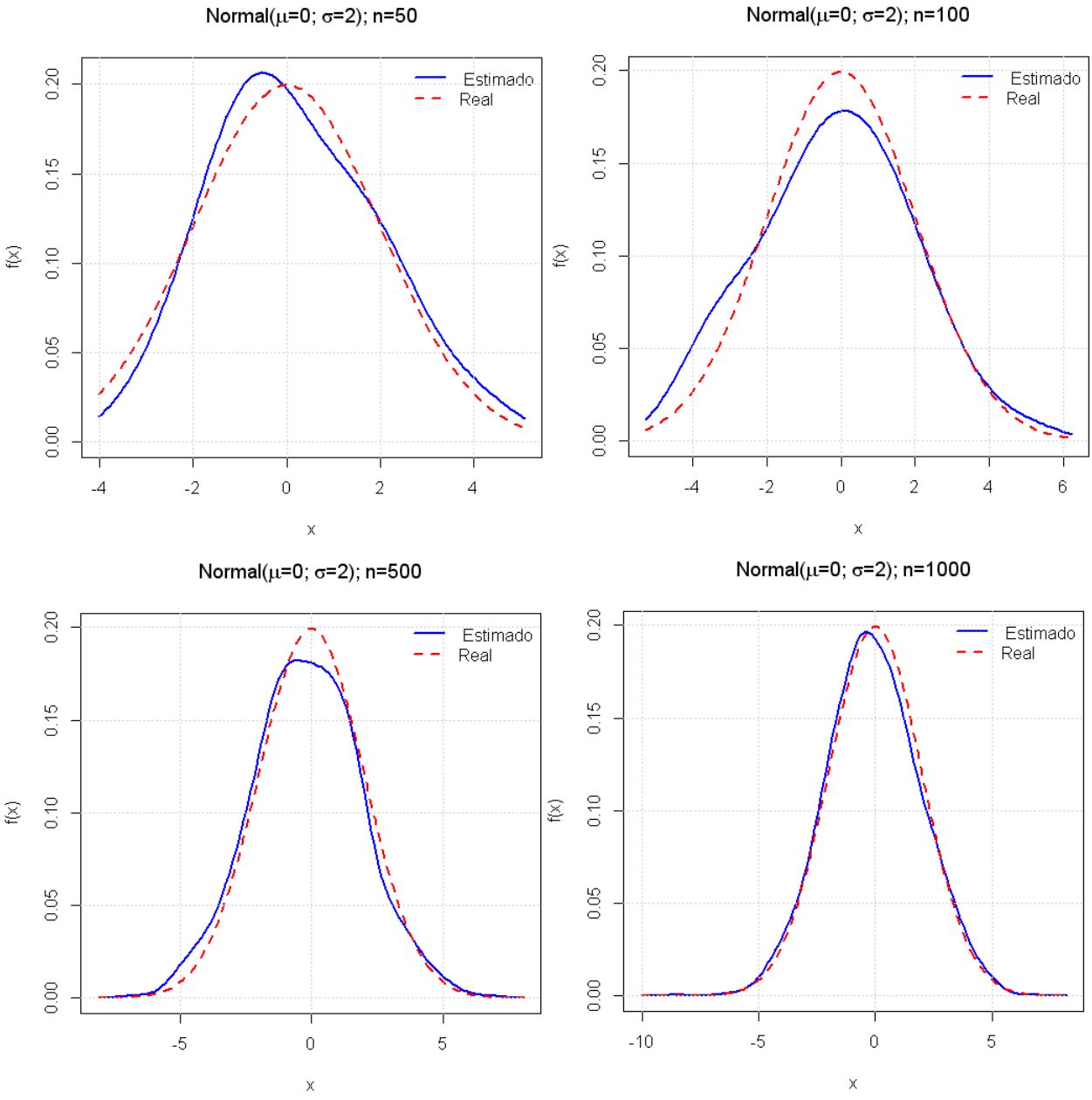

Figura 7: Comparação entre a função de densidade estimada via núcleo estimador (—) e a função de densidade real (- - -) para uma Normal (0,4).

2.7.2 Distribuição Weibull (2,3)

No *software* "R" geramos amostras da distribuição de Weibull com parâmetro de forma $\alpha = 2$ e parâmetro de escala $\beta = 3$ para $n = \{50, 100, 500, 1000\}$. A densidade foi estimada através das equações (2.1) e (2.2) e comparamos as estimativas com a função real, veja *Figura 8*:

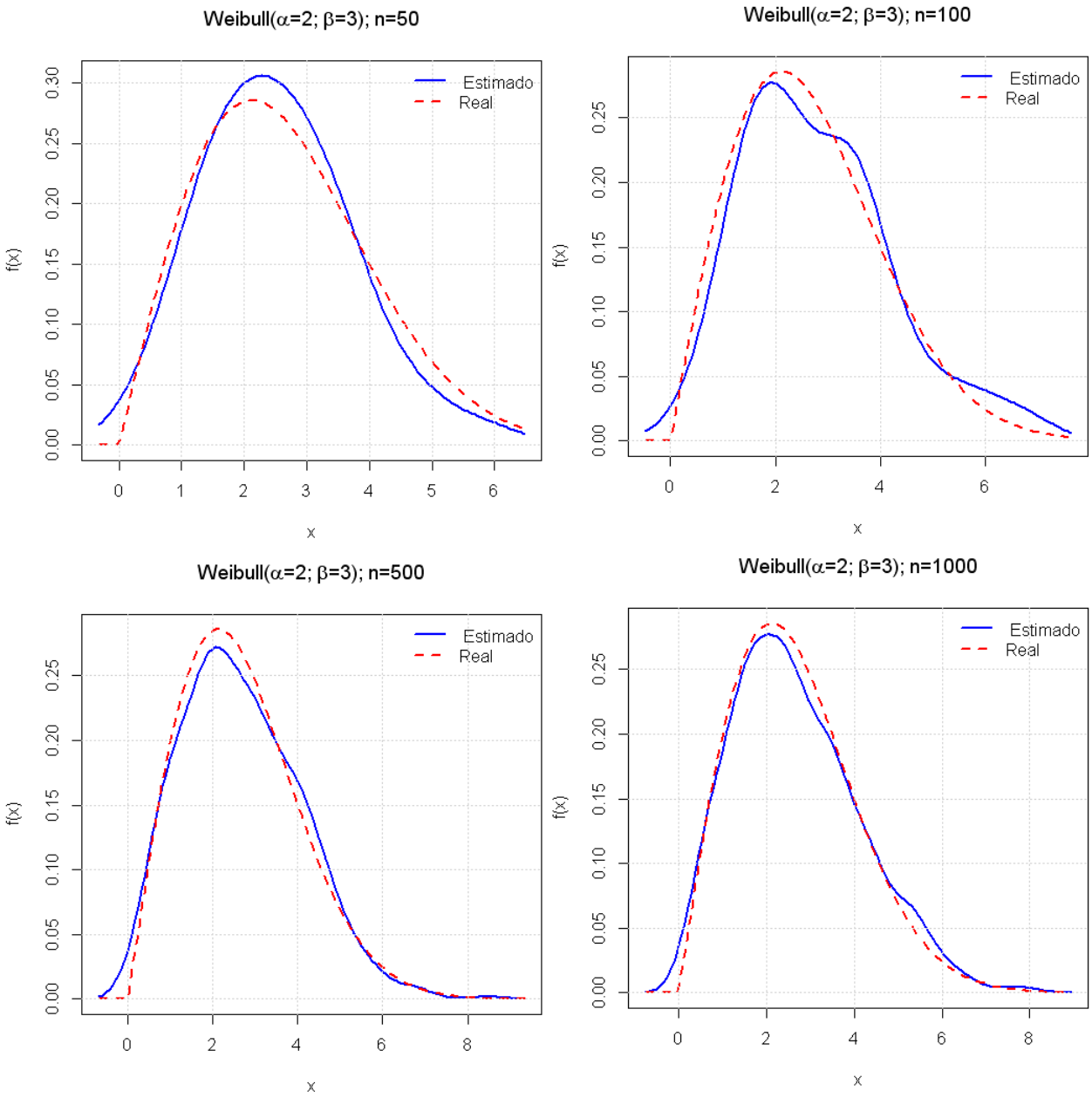

Figura 8: Comparação entre a função de densidade estimada via núcleo estimador (—) e a função de densidade real (- - -) para uma Weibull (2,3).

Podemos verificar que o estimador, aproxima-se bem da função de densidade, principalmente quando “n $\geq$ 50”. Podemos concluir que quanto maior o valor de “n” melhor será a aproximação do estimador à função real.

3 Núcleo-estimador da Função de Distribuição.

Neste capítulo definiremos a função de distribuição empírica, o Núcleo-estimador para a função de distribuição, a janela ótima para este estimador e, como realizado no capítulo anterior, faremos uma comparação gráfica entre o Núcleo-estimador e a função de distribuição de amostras de dados normais e wiebull.

3.1 A Função de Distribuição Empírica

A função de distribuição empírica é um dos estimadores mais utilizados quando se deseja estimar a função de distribuição F de uma variável aleatória (SILVERMAN, 1986).

Considere uma amostra aleatória de tamanho n denotada por, $X_1; X_2; \ldots; X_n$, da variável aleatória X, que tem função de distribuição acumulada F. Um estimador F, avaliado em x é dado por:

$$\tilde{F}_n(x) = P(X \leq x) = \frac{\text{número de valores observados no intervalo } (-\infty, x]}{n} \tag{3.1}$$

A função de distribuição empírica é uma representação gráfica que nos fornece a informação do tipo da percentagem de valores da amostra inferiores ou ou superiores a um determinado quantil.

Entretanto, $\tilde{F}_n$ algumas vezes apresenta problemas na estimativa das ordenadas correspondentes aos quantis. Isto é explicado pelo fato de que esta seja uma função escada que pondera cada x_i amostral pelo valor $\frac{1}{n}$. Assim quando o tamanho da amostra é pequeno, os valores estimados muitas vezes não se aproximam dos valores reais da função teórica F, ou seja quanto menor o tamanho amostral "n" maior será o "degrau" desta função escada. Como exemplo verifique a *Figura 9*.

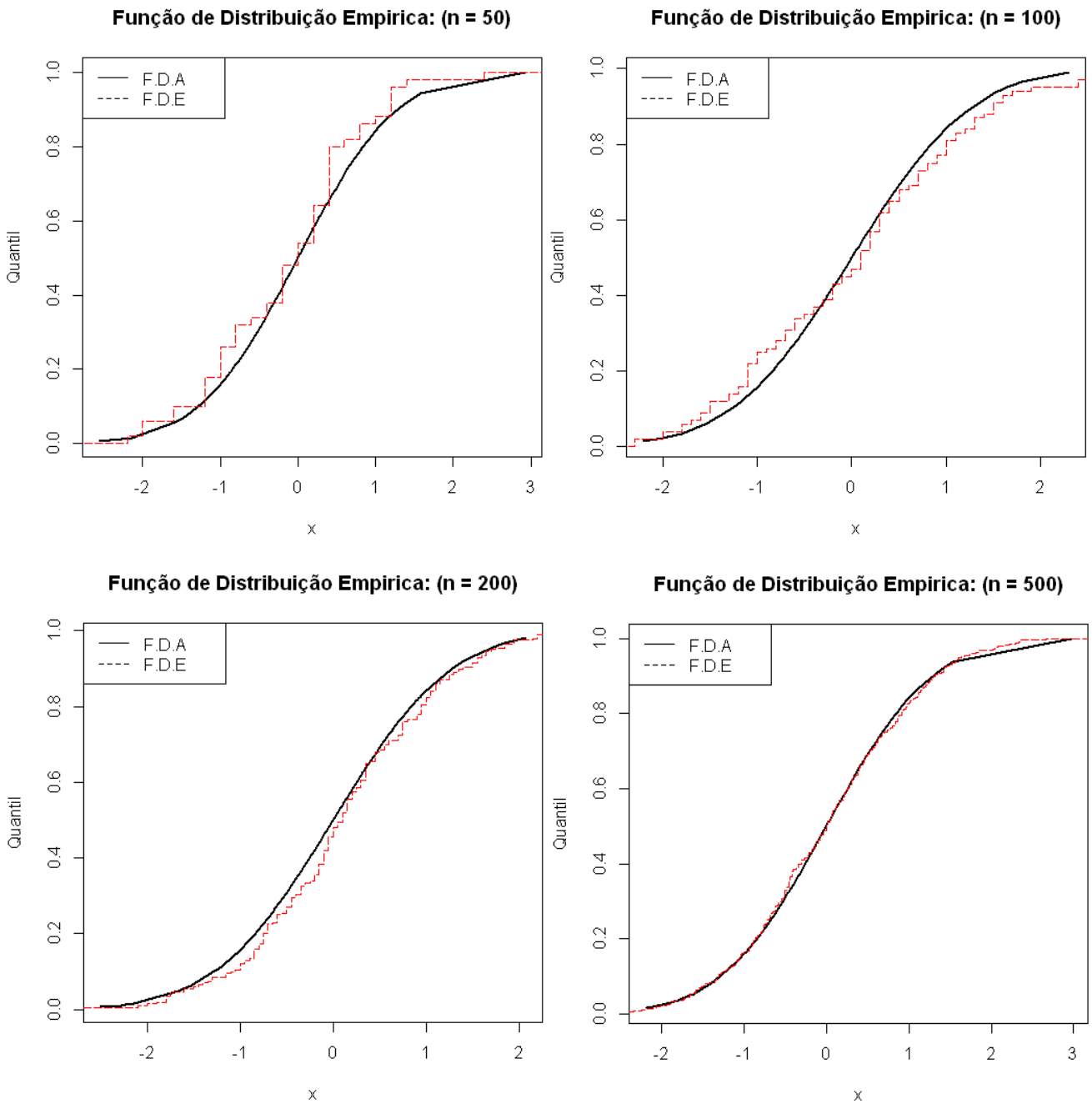

Figura 9: Comparação entre a função de distribuição empírica F.D.E (- - -) e a função de distribuição acumulada F.D.A (—) para uma amostras de tamanho n=$\{50, 100, 200, 500\}$ proveniente de uma distribuição Normal (0,1)

3.2 Núcleo Estimador da Função de Distribuição

O método do núcleo estimador também tem sido amplamente utilizado na estimação da função de distribuição, consolidando-se como uma alternativa às abordagens paramétricas.

Dada uma amostra aleatória, $X_1; \cdots ; X_n$, de uma variável aleatória contínua X, com função de distribuição F, define-se o núcleo estimador de F, avaliado no ponto x, por:

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h_n}\right) \tag{3.2}$$

onde K é uma função de distribuição denominada "núcleo" e h_n chamado de parâmetro de suavidade ou janela ótima para a função de distribuição, que será discutido na próxima subseção.

3.2.1 Janela Òtima da Função de Distribuição

Assim como tinhamos no caso da função de densidade, a escolha do núcleo K não é uma tarefa muito difícil. Mas a escolha da janela h_n é um sério problema, que vem sendo tratado exaustivamente na literatura. O parâmetro de suaviadade, h_n, é aquele que minimiza a diferença entre as estimativas obtidas através do núcleo estimador $\hat{f}(x)$ e os verdadeiros valores da função densidade $f(x)$, ou seja, queremos a janela minimize o Erro Quadratico Integrado Médio ($EQIM$), sendo definido por:

$$EQIM(h) = E \int \left\{\hat{F}_n(x) - F(x)\right\}^2 dx \tag{3.3}$$

Como no caso da função de densidade, existe uma expressão para a janela ótima que minimiza o EQIM(h). Mas como visto anteriormente h_{opt} infelizmente, depende da função desconhecida F. Precisamos então, estimar h_{opt} a partir dos dados observados. Em (BOWMAN; HALL; PRVAN, 1998) obtemos a expressão da janela ótima dada por:

$$h_{opt} = \left\{ \frac{\int W(x)\left[1 - W(x)\right] dx}{\left[\int z^2 dW(z)\right]^2 \int \left[F''(x)\right]^2 dx} \right\}^{\frac{1}{3}} n^{-\frac{1}{3}} \tag{3.4}$$

onde $W(x)$ é a função de distribuição normal (0,1) no ponto x e $H = \int |F''(x)|^2 dx$, ou por definição, $\hat{H} = \int |f'(x)|^2 dx$, podendo, como no caso da sensidade, ser estimado

por $\hat{H}$ obtido pelo método plug-in através da equação:

$$\hat{H} = \frac{1}{\pi} \int_{o}^{\Lambda} \lambda^2 \left[|\hat{\varphi}(\lambda)|^2 - \frac{1}{n} \right] d\lambda \tag{3.5}$$

onde: $\hat{\varphi}(\lambda)$ é a função característica empírica, $\Lambda = min\left\{\lambda : |\hat{\varphi}|^2 \leq \frac{c}{m}\right\}$

Substituindo $\int [F''(x)]^2 dx$ por $\hat{H}$ na equação (3.4), temos:

$$\hat{h}_{opt} = \left\{ \frac{\int W(x)\,[1 - W(x)]\,dx}{\left[\int z^2 dW(z)\right]^2 \hat{H}} \right\}^{\frac{1}{3}} n^{-\frac{1}{3}} \tag{3.6}$$

Para maiores esclarecimentos ver cap. 3 de (BESSEGATO; ATUNCAR; DUCZMAL, 2001)

3.3 Comparação Gráfica - Função de Distribuição

No *software* "R" geramos amostras com distribuição normal com média 10 e desvio padrão igual à 3,5; para $n = \{50; 100; 500; 5000\}$ e estimamos a função de distribuição através das equações (3.2) e (3.6), comparando-os com sua função-distribuição real. Verificamos que quanto maior é o valor de "n", menor será o resíduo do estimador, ou seja, quanto maior for o valor de "'n" mais próximo o estimador estará da função-distribução real, Como mostra a *Figura 11.*

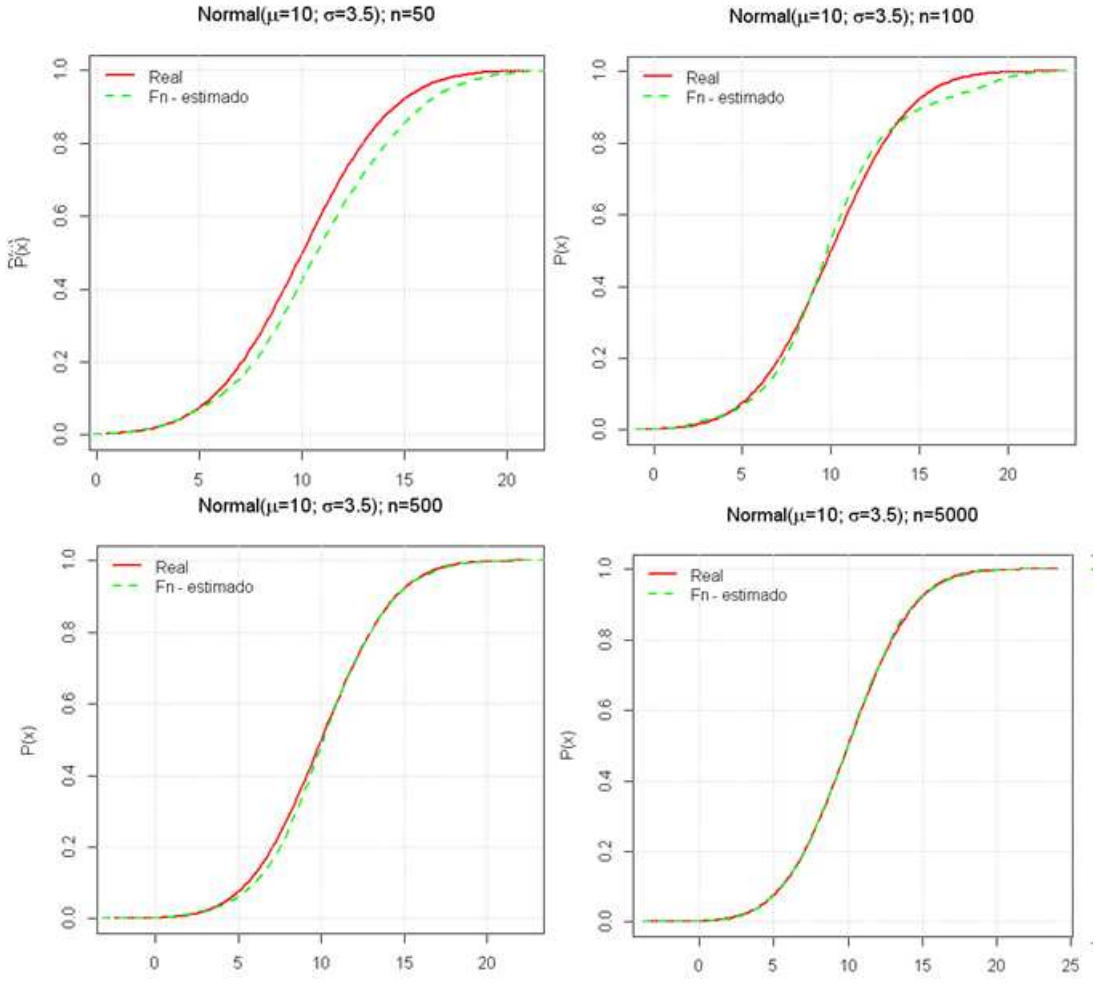

Figura 10: Comparação entre a função de distribuição estimada via núcleo estimador e a função de distribuição real para uma normal com $\mu = 10$ e $\sigma = 3,5$.

3.4 Comparação Gráfica - Função Característica.

Nesta seção comparamos a função característica teórica $\varphi(\lambda)$ com a função característica empírica $\hat{\varphi}(\lambda)$, dada pelas equações (2.11) e (2.12).

Analisamos o ajuste da função característica para vários valores de "n". Esta comparação foi feita gerando amostras normais de tamanho $n = \{30; 50; 100; 500; 1000; 5000\}$ para uma sequência de $\lambda = \{0.1, 0, 2; 0, 3; \cdots ; 4, 8; 4, 9; 5.0\}$.

Então, de (2.14), definimos o número complexo $z = a_\lambda + i.b_\lambda$ onde:

$$a_\lambda = \frac{1}{n}\sum_{j=1}^{n} cos\,(\lambda X_j)$$

$$b_\lambda = \frac{i}{n}\sum_{j=1}^{n} sen\,(\lambda X_j)$$

Assim podemos definir o módulo da função característica empírica como:

$$|\hat{\varphi}_x(\lambda)|^2 = a_\lambda^2 + b_\lambda^2 \tag{3.7}$$

As comparações foram realizadas entre o o quadrado do módulo da função característca empírica dada pela equação (3.7), "linha pontilhada" e o quadrado da função característica da normal padrão, dada por, $\varphi_x^2(\lambda) = \left(e^{-\frac{1}{2}\lambda^2}\right)^2$ "linha sólida". Nos gráficos apresentados na *Figura 10* podemos verificar que a função $|\hat{\varphi}_x(\lambda)|^2$ se ajusta bem à função $(\varphi_x(\lambda))^2$ quando "$n \geq 50$". Observamos também que quanto maior o tamanho amostral mais próximas entre si, estão estas funções.

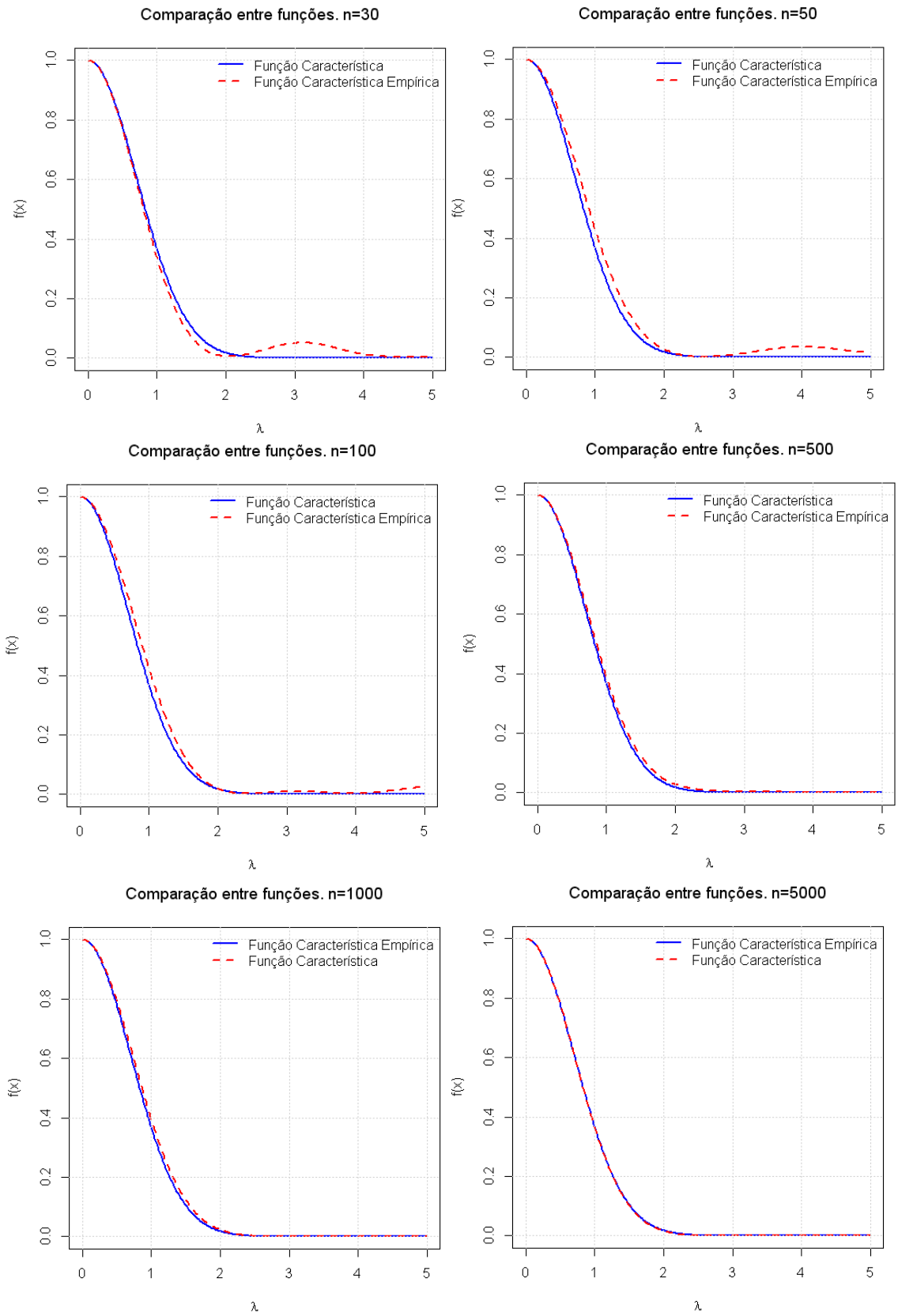

Figura 11: Comparação entre as funções de distribuição empírica (amostral e teórica) em função de λ, para $n = \{30, 50, 100, 500, 1000, 5000\}$

Sabemos que em situações reais o tamanho amostral raramente é superior à 100. Nossa opção por simularmos amostras com “n” iguais à 500, 1000, e 5000 tem como objetivo verificar que quanto maior o valor de “n” melhor será o ajuste entre as funções caracterísica e a estimada.

4 Aplicação.

4.1 A Transformação Linear

A técnica de núcleo estimador aplica-se quando desejamos estimar a função de densidade, ou distribuição, de um conjunto de dados que não se ajusta às funções paramétricas conhecidas. Entretanto, como não temos nenhum controle sobre este conjunto de dados, podemos ter a variância amostral muito pequena ou extremamente grande. Como visto anteriormente, usando o método estudado para estimar a janela, este estimador só pode ser calculado para funções "suaves" o que nem sempre pode ocorrer.

Para estimarmos estas funções devemos verificar a magnitude de sua variância amostral. Quando estas variâncias são muito grandes ou muito pequenas o cálculo de $\Lambda = min\left\{\lambda : |\hat{\varphi}|^2 \leq \frac{c}{m}\right\}$ através da função característica, fica comprometido. Esta dificuldade, como visto anteriormente, deve-se à demora na convergência do $|\varphi|^2$ no caso onde a variância é muito pequena ou no decaímento muito rápido desta função, quando temos variância muito grande.

Devido a este problema, é recomendávfel aplicar uma transformação linear nos dados amostrais. Esta transformação consiste em dividir cada elemento amostral x_i pelo desvio padrão amostral s criando assim uma nova amostra que denominaremos por $\{y_1, y_2, \cdots, y_n\}$. Este nova amostra será transformada na amostra original x_i após calcularmos o núcleo estimador através das equações (2.1) ou (3.2), ou seja, voltará à função original após estimarmos sua função transformada.

Para a função de distribuição F_x, sendo $x = s.y$, esta transformação é dada por:

$$F_X(x) = P(X \leq x) = P(s.Y \leq x) = P\left(Y \leq \frac{x}{s}\right) = F_Y\left(\frac{x}{s}\right)$$

temos então que:

$$F_X(x) = F_Y\left(\frac{x}{s}\right) \tag{4.1}$$

Para encontrarmos a função de densidade, derivamos a equação (4.1), em ambos os lados, em relação à x, temos:

$$\frac{\partial\left(F_X(x)\right)}{\partial x} = f_x(x) \tag{4.2}$$

$$\frac{\partial\left(F_Y\ \left(\frac{x}{s}\right)\right)}{\partial x} = \frac{1}{s}.f_y\left(\frac{x}{s}\right) \tag{4.3}$$

igualando as equações (4.2) e (4.3) temos:

$$f_x(x) = \frac{1}{s}.f_y\left(\frac{x}{s}\right) \tag{4.4}$$

Assim a partir de (4.1) e (4.4) definimos:

$$\hat{F}_X(x) = F_Y\left(\frac{x}{s}\right) e \hat{f}_x(x) = \frac{1}{s}.f_y\left(\frac{x}{s}\right)$$

que é a transformação inversa feita em (4.1) para a função de densidade.

4.1.1 Exemplo de Aplicação

Suponha que tenhamos o conjunto com 5 elementos dado por $a_i = \{14,8; 50,6; 70,3; 80,4; 93,6\}$, que tem como desvio padrão amostral igual à $s = 29,85$ e média amostral $\bar{x} = 62,28$.

Iniciaremos a transformação definindo os 100 pontos da malha que representaram a abcissa, os quais chamaremos de x_i sendo $i = (1; 2; \cdots; 100)$. O primeiro e o último ponto desta malha são dados por:

$$x_{(1)} = a_{min} - h_{opt}; \cdots; x_{(100)} = a_{max} + h_{opt}$$

onde h_{opt} é a janela ótima do conjunto de dados $y_i = \frac{a_i}{s}$. O tamanho desta janela ótima foi calculada através da equação (2.16) para a amostra y_i, tendo como resultado $h_{opt} = 1,4249$. Assim o primeiro e último termos da malha são dados por:

$$x_{(1)} = 13,3754; \cdots; x_{(100)} = 95,0249$$

Temos mais 98 pontos para serem distribuídos no interior desta malha. Estes pontos serão calculados usando "interpolação aritmética", que tem sua razão aritmética dada pela fórmula:

$$R = \frac{\left(x_{(n)} - x_{(1)}\right)}{(n-1)} \tag{4.5}$$

Para este caso, temos "n=100" e calculando-se a razão desta interpolação, através da fórmula acima, temos que $R = 1,669$. Assim, os demais pontos $x_{(2;\cdots;99)}$ são dados por:

$$x_{(1)} = 13,3754; x_{(2)} = 14,1998; x_{(3)} = 15,0246; \cdots ; x_{(99)} = 94,2002; x_{(100)} = 95,0249$$

que são os 100 pontos da malha ainda não transformados.

Agora iremos transformar a malha x_i, ou seja, iremos definir um novo conjunto $v_i = \frac{x_i}{s}$, sendo que este conjunto, transformado, tem variância igual à 1 e não apresenta problemas no cálculo do Λ.

$$v_{(1)} = \frac{x_{(1)}}{s} \Rightarrow v_{(1)} = \frac{13,3754}{29,85} = 0,4367$$
$$v_{(2)} = \frac{x_{(2)}}{s} \Rightarrow v_{(2)} = \frac{14,1998}{29,85} = 0,4636$$
$$\vdots$$
$$v_{(99)} = \frac{x_{(99)}}{s} \Rightarrow v_{(99)} = \frac{94,2001}{29,85} = 3,0754$$
$$v_{(100)} = \frac{x_{(100)}}{s} \Rightarrow v_{(100)} = \frac{95,0249}{29,85} = 3,1024$$

Como feito acima, iremos transformar também os dados amostrais a_i, definindo um novo conjunto $y_i = \frac{a_i}{s}$, sendo que este conjunto também não apresenta problemas no cálculo do Λ.

$$y_{(1)} = \frac{a_{(1)}}{s} \Rightarrow y_{(1)} = \frac{14,8}{29,85} = 0,496$$
$$y_{(2)} = \frac{a_{(2)}}{s} \Rightarrow y_{(2)} = \frac{50,6}{29,85} = 1,697$$
$$\vdots$$
$$y_{(5)} = \frac{a_{(5)}}{s} \Rightarrow y_{(5)} = \frac{93,6}{29,85} = 3,136$$

Após estas transformações, com a variância controlada, aplicaremos a fórmula (2.1) para estimarmos a função de densidade ou a fórmula (3.2) para estimarmos a função de distribuição.

4.1.2 Estratégia de Implementação

Neste exemplo de aplicação temos como "x_i" e "X_j" os elementos dos conjuntos "v_i" e "y_i" respectivamente, assim a fórmula (2.1) fica representada por:

$$f_n(v) = \frac{1}{nh_n} \sum_{i=1}^{n} k\left(\frac{v_i - y_j}{h_n}\right)$$

Onde $k\,(x)$ representa a função de densidade da normal padrão no ponto "x" e h_n é a janela ótima da amostra dada pela equação (2.2).

Neste exemplo de aplicação temos que "n=100" e "$h_n = h_{opt} = 1,4249$". Assim a função de densidade do exemplo de aplicação, será calculada de seguinte forma:

$$f_1(v) = \frac{1}{nh_n}\left[k\left(\frac{v_1 - y_1}{h_n}\right) + k\left(\frac{v_1 - y_2}{h_n}\right) + \cdots + k\left(\frac{v_1 - y_5}{h_n}\right)\right] = 0,20026$$

$$f_2(v) = \frac{1}{nh_n}\left[k\left(\frac{v_2 - y_1}{h_n}\right) + k\left(\frac{v_2 - y_2}{h_n}\right) + \cdots + k\left(\frac{v_2 - y_5}{h_n}\right)\right] = 0,20181$$

$$\vdots$$

$$f_{100}(v) = \frac{1}{nh_n}\left[k\left(\frac{v_{100} - y_1}{h_n}\right) + k\left(\frac{v_{100} - y_2}{h_n}\right) + \cdots + k\left(\frac{v_{100} - y_5}{h_n}\right)\right] = 0,30948$$

Assim após a estimativa da densidade transformada, utilizamos a equação (4.4) para retornar à variável original. Ainda, com esta transformação, podemos calcular a função de distribuição dos dados através da fórmula (3.2), tendo como janela ótima h_n o valor resultante da aplicação da equação (3.4).

4.1.3 Janela Ótima na transformação

Quando usamos a transformação descrita na seção (4.1.1), a janela ótima inicial, utilizada na malha, é calculada através dos dados transformados y_i. Já a janela que utilizamos para o cálculo do Núcleo-estimador é obtida através do conjunto $v_i = \frac{x_i}{s}$. Em nosso al-

goritmo utilizamos como critério para a transformação linear os desvios amostrais que se encontram fora do intervalo $0,5 \leq \hat{s} \leq 3$.

Nosso algoritmo utiliza 100 repetições, ou seja gera 100 amostras com média μ e desvio padrão σ. Quando este desvio está próximo da região limítrofe acima, podemos ter em algumas destas repetições, desvios amostrais s que necessitam de transformação e outros que não necessitam de tal transformação para o mesmo σ^2. Quando isto ocorre, verifica-se as janelas ótima se separam, claramente, em dois grupos distintos; o grupo onde foi feita a transformação e o grupo onde não foi feita a transformação. Veja as *Figuras 12 e 13* a seguir.

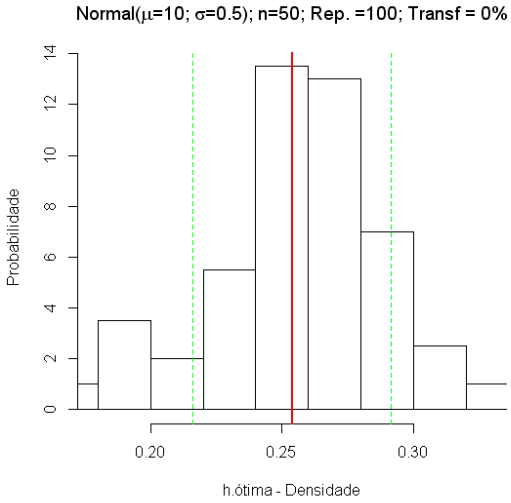

Figura 12: Histograma para a janela ótima com 0% de transformações.(Apenas um grupo).

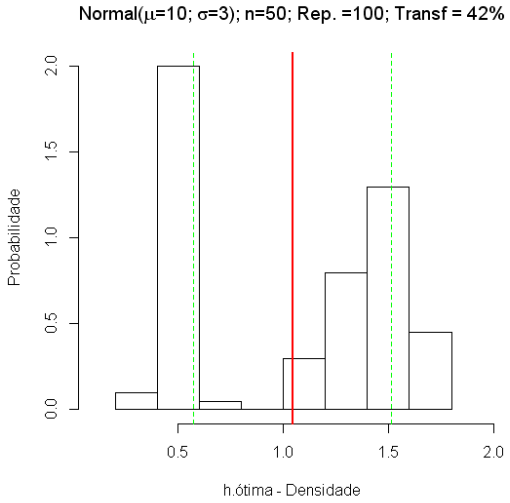

Figura 13: Histograma para a janela ótima com 42% de transformações.(Dois grupos distintos).

Nosso algoritmo gerou simulações normais com média 10, tamanho amostral $n = \{30, 50, 100, 500\}$ e com desvios $\sigma = \{0.1, 0.5, 2, 3, 3.5, 5\}$ gerando um total de 20 histogramas a serem analisados. Podemos verificar qua quando "n" aumenta, o tamanho da janela e a variância diminuem, ou seja, se $n \to \infty$ então $h_n \to 0$, $s_{\hat{h}} \to 0$ e $n.h_n \to \infty$. Este comportamento foi observado para todos os valores de σ estudados.

A título de ilustração, colocaremos os gráficos para $\sigma = 2$. Conforme *Figura 14*.

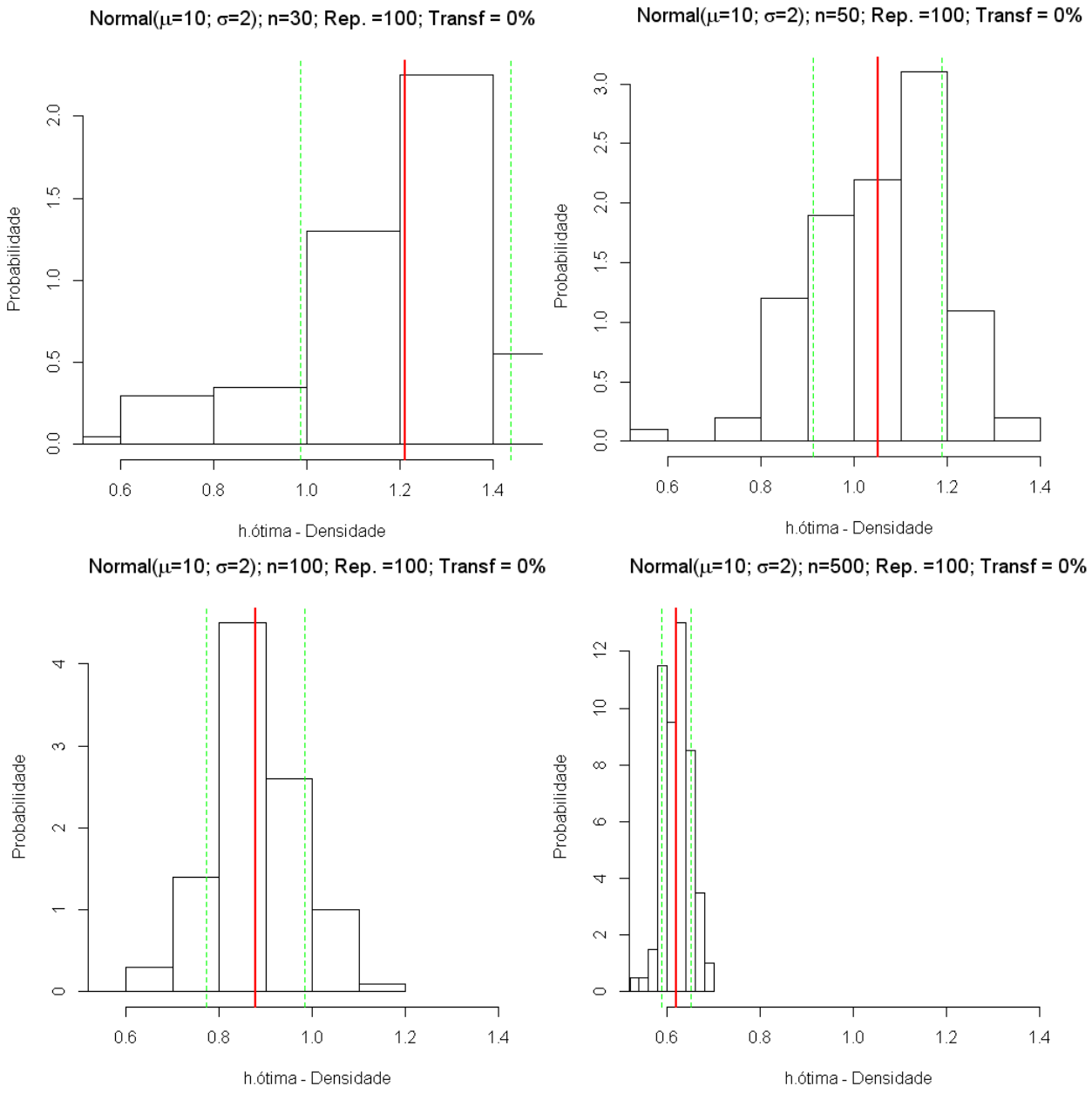

Figura 14: Histogramas para a Janela Ótima para $\sigma = 2$ com 0% em transformações.

4.2 Aplicação à base de dados

Nesta seção tentaremos ajustar à base de dados T-anos várias densidades conhecidas. Caso estes dados não se ajustem à nenhuma destas densidades, iremos estimar sua função densidade utilizando Núcleo-estimador. Tentaremos ajustar estes dados através do *software Minitab for Windows* versão 15, em sua opção $(Graph) \rightarrow (Probability - Plot)$.

4.2.1 A base de dados T-anos

A base de dados "T-anos" refere-se ao tempo de duração, em anos, de 62 componentes eletrônicos de uma certa empresa. O conjuto de dados, T-anos é representado por:

T-anos =	(3,4922;	0,6077;	8,3922;	3,9138;	4,3348;	4,1701;	4,2057;	4,0198;
7,2889;	1,2353;	6,8360;	4,0694;	12,465;	7,3819;	4,8963;	9,1073;	4,0694;
2,4193;	7,4869;	5,0387;	0,8399;	1,4097;	4,0041;	8,0956;	6,7759;	4,2472;
3,6368;	0,6706;	9,8407;	6,5755;	1,2131;	3,6599;	2,9506;	2,0630;	0,4476;
4,2472;	1,5879;	4,0504;	3,2648;	1,2778;	4,2893;	5,3322;	1,8517;	3,0793;
5,0387;	3,3469;	4,2608;	3,1828;	2,4699;	0,8993;	6,1891;	8,8372;	4,7425;
7,4869;	4,2901;	3,8879;	9,0383;	8,0956;	2,4699;	6,3950;	5,4196;	6,7759)

Com esta amostra, tentaremos ajustá-la às distribuições Normal, lognormal, logística e Weibull. Para iniciarmos a análise destes dados faremos uma análise descritíva, através do *software Minitab* desta amostra, tendo como resultado:

Variable	Mean	StDev	Q1	Median	Q3	Skewness	Kurtosis
Tanos	4,575	2,617	2,830	4,188	6,626	0,60	0,15

Podemos verificar uma concentração maior de massa à esquerda da média (*assimetria*), devido à *skewness* positiva, além de um pico um pouco mais acentuado que a normal devido à *kurtosis* positiva.

Para termos uma idéia da forma da função de densidade desta amostra mostraremos um histograma destes dados na *Figura 15*, que confirma a análise descritiva dos dados.

Figura 15: Histograma para a base de dado T-anos.

4.3 Teste de Aderência

O teste não paramétrico, utilizado neste *software*, para testar a aderência da distribuição à amostra é o teste de Anderson-Darling (AD). Este é um teste unicaudal e tem sua hipótese nula rejeitada quando a estatística de teste, que depende da função de distribuição acumulada e do tamanho amostral "n", supera um valor crítico previamente calculado. Para maiores informações veja (ANDERSON; DARLING, 1954).

Realizamos este teste, sob a base de dados T-anos, para as distribuições normal, log-normal, logística e weibull. Os resultados encontram-se nas *Figuras 16 e 17*.

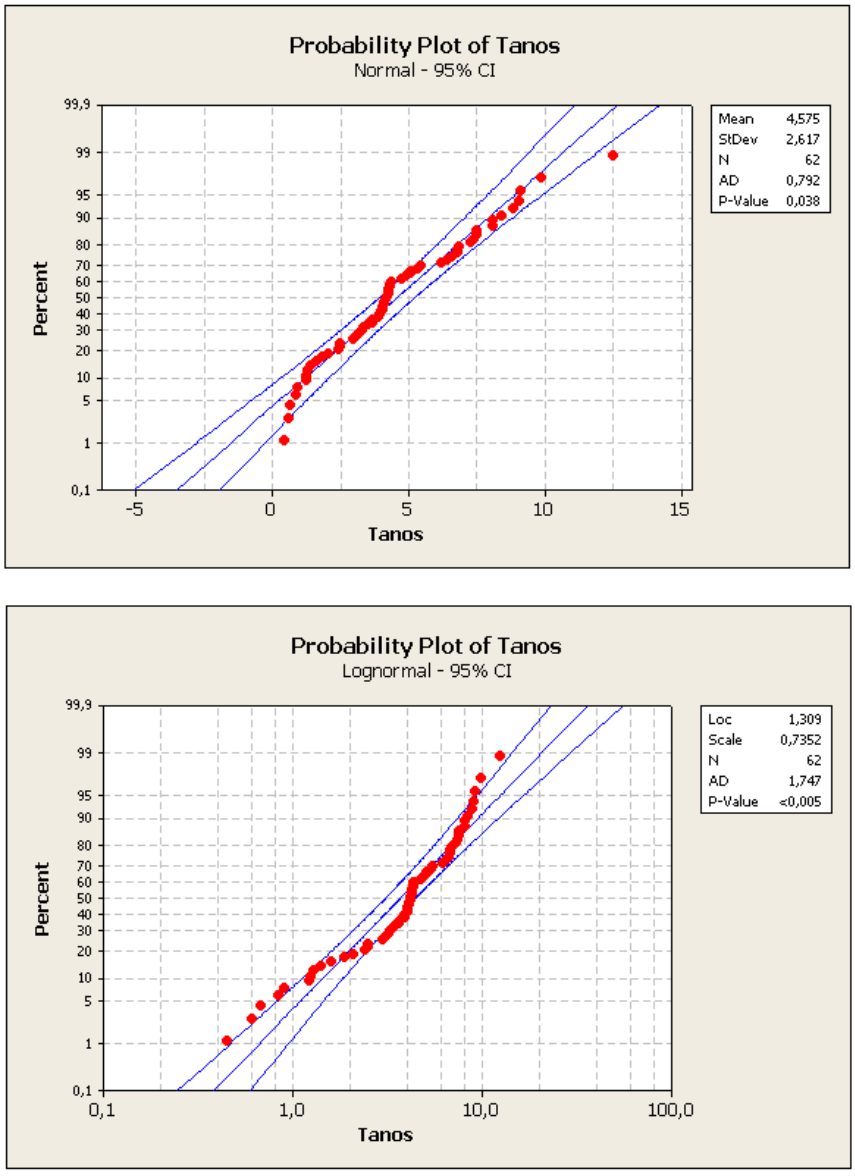

Figura 16: QQplot - Testando a aderência da base de dados à distribuição normal e log-normal.

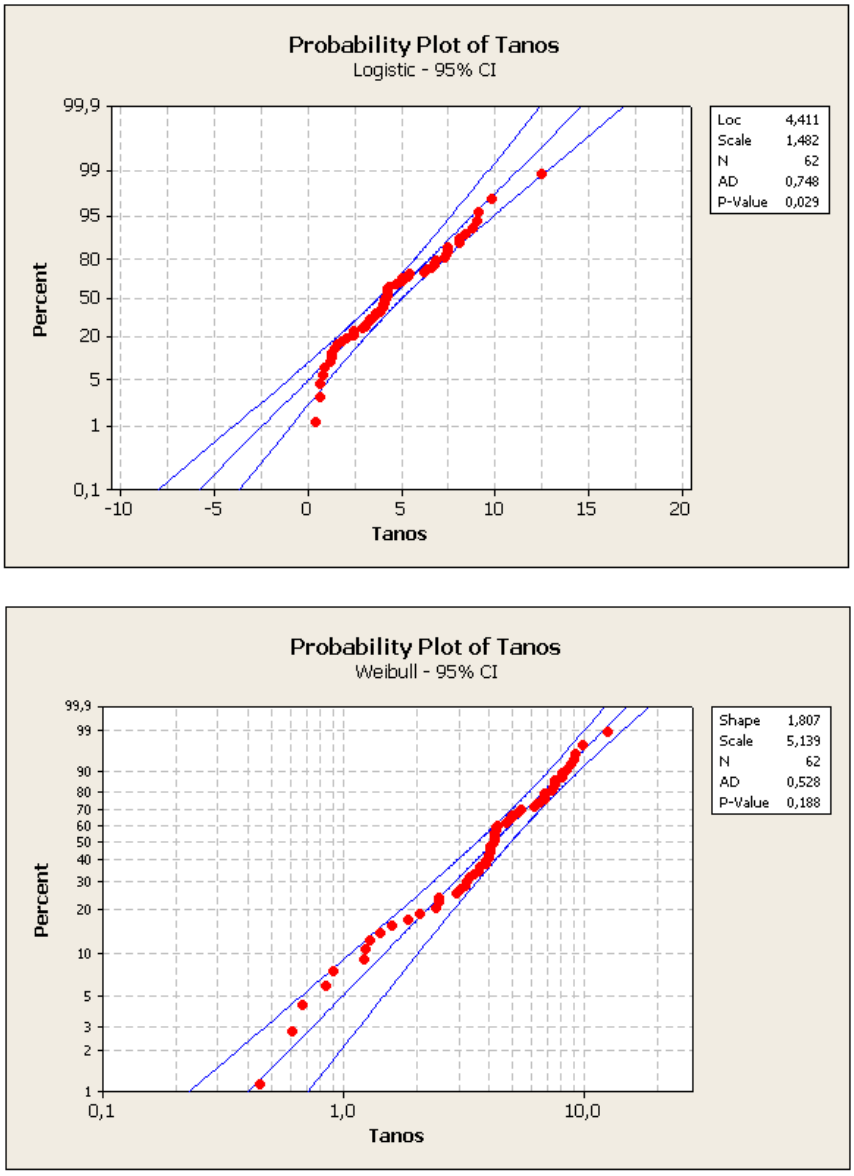

Figura 17: QQplot - Testando a aderência da base de dados à distribuição logística e Weibull.

Na *Tabela 1* comparamos os p-valores gerados a partir do teste de aderência (AD) para estas densidades.

DISTRIBUIÇÃO	**p-valor**
normal	0,038*
log-normal	<0,005*
logística	0,029*
weibull	0,188

Tabela 1: P-valor para o teste de aderência.

Podemos verificar que as distribuições normal, log-normal e logística apresentam p-valores ($\leq 0,05*$), ou seja, à um nível de significância de esta amostra não se ajusta à nenhuma destas distribuições. Já a distriuição de Weibull apresenta p-valor de 0,188 que pode ser considerado um p-valor significativo ao nível de 5%. Na *Figura 18* mostramos o ajuste das quatro densidades ao histograma da amostra e verificamos que, realmente, apenas a distribuição de Weibull se ajusta melhor ao histograma.

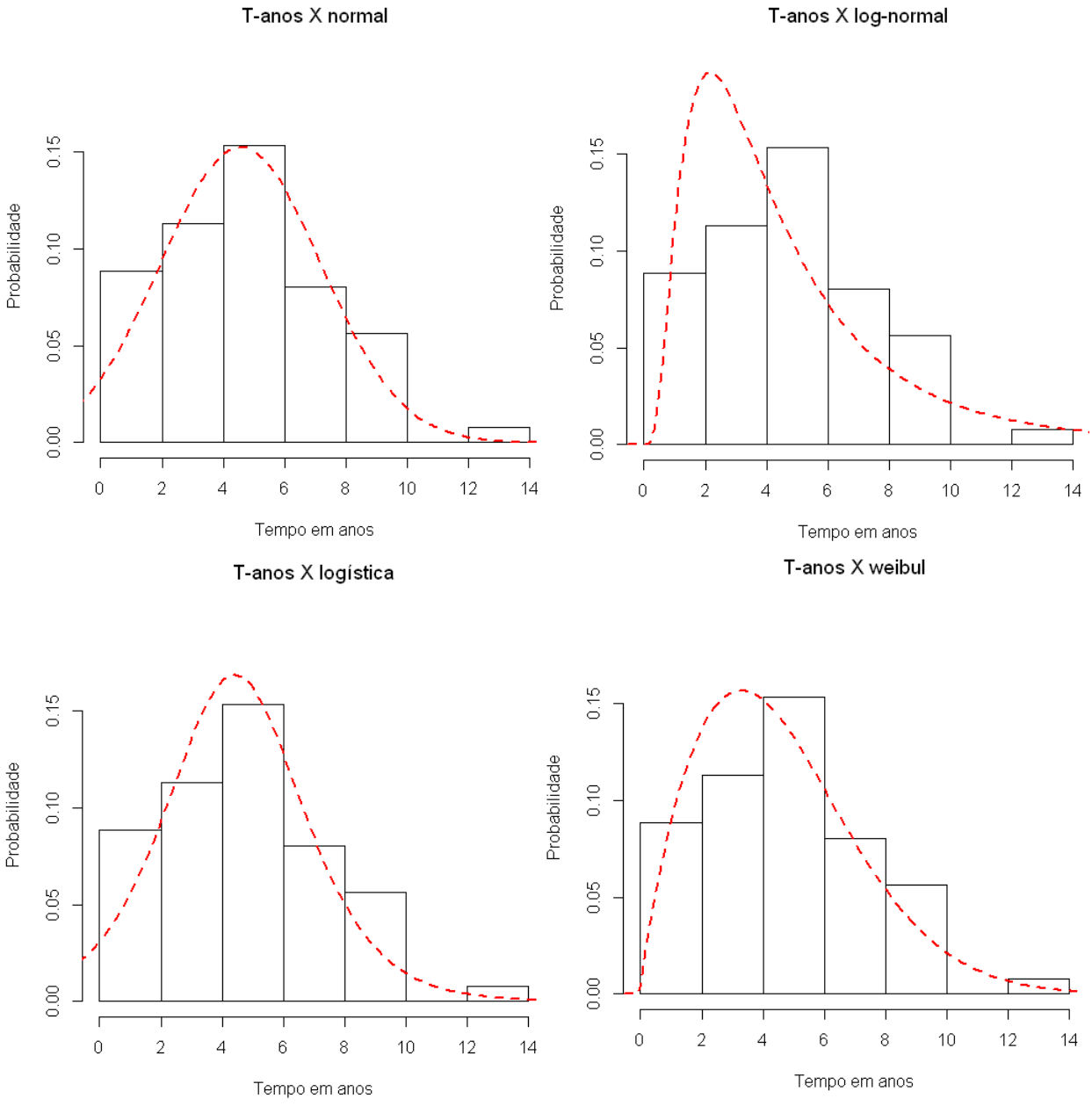

Figura 18: Comparação entre as densidades e o histograma da amostra.

4.3.1 Comparação via Núcleo-estimador

Com a base de dados "t-anos", testamos várias distribuições paramétricas de probabilidade, sendo que apenas a distribuição de Weibull se ajustou, ao nível de 5%, ao conjunto de dados. Esta distribuição foi a que gerou o maior p-valor (0.188).

Nesta seção iremos comparar graficamente a técnica, não paramétrica, de Núcleo-estimador com a distribuição de Weibull, que foi significativa, ao ajuste dos dados. Podemos verificar, através da *Figura 19* que o Núcleo-estimador, neste caso, gerou menor resíduo que a distribuição de Weibull, quando comparado com o histograma.

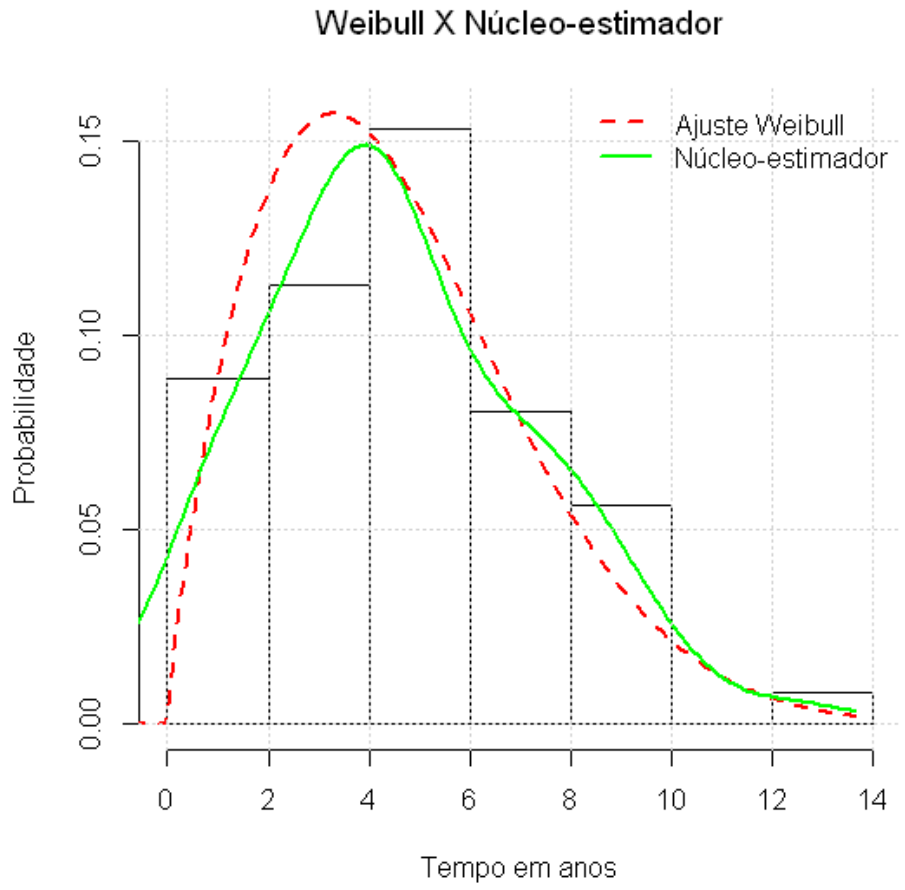

Figura 19: Comparaçao entre o Núcleo-estimador e a densidade de Weibull.

4.4 Conclusões e trabalhos futuros

Os métodos não paramétricos de estimação funcional são muito úteis para obtermos informações sobre um conjunto de dados cuja distribuição não pode ser aproximada por modelos paramétricos.

Neste trabalho analisamos a importância no cálculo da "janela ótima" para encontrar um bom estimador, e posteriormente fazermos uma análise mais detalhada do estimador funcional. Os resultados da janela ótima, "h_{opt}" e os resultados dos estimadores "$\hat{H}$" para a densidade e "$\hat{G}$" para a distribuição, foram calculados, definidos, provados e implementados por (BESSEGATO, 2001) através do método *Plug-in* modificado. A transformação linear, as comparações gráficas e os gráficos para análise da janela ótima foram implementados aqui, já os resultados prontos, apresentados aqui, para a janela ótima e as estimativas $\hat{G}$ e $\hat{H}$, foram calculados através do método *Plug-in* e retirados do código do Lupércio.

Verificamos que o Núcleo-estimador é uma boa técnica para estimar funções de densidade ou distribuição não paramétricas que sejam suaves e com amostras de tamanho ≥ 30. Neste trabalho verificamos, na fase de simulação, que quando aumentamos o tamanho amostral "n" menor será o resíduo do estimador.

Na parte computacional, optamos em implementar os algoritmos e códigos no *software R*, versão 3.0.3 (R Core Team, 2014), devido ao seu livre acesso e pela facilidade em implementar códigos no mesmo. Tivemos algumas dificuldades em implementar o algoritmo referente à transformação linear, mas após algum tempo, e várias tentativas, o código funcionou como esperado. Entretanto, para os cálculos onde tinhamos várias repetições, como na subseção "*Janela ótima na transformação*", verificamos que o tempo gasto para o programa calcular os resultados foi muito alto, principalmente para resultados onde tinhamos tamanhos amostrais "$n \geq 100$".

Entretanto, trabalhamos apenas com janelas "globais", tamanho fixo, mas na literatura podemos encontrar várias outras técnicas que trabalham com janela de tamanho variável ou com problemas de fronterias. Acreditamos que uma extensão dos tópicos estudados aqui, para janelas variáveis, seria aplicação relevante para a continuação deste trabalho.

Referências

R Core Team. [S.l.: s.n.].

ANDERSON, T. W.; DARLING, A. D. An test of goodness of fit. *Journal of the American Statistical Association*, v. 49, n. 268, p. 765–769, 1954.

ATUNCAR, G. S.; DAMASCENO, E. C.; MENDONçA, P. P. Choosing the bandwidth in nonparametric functional estimation. *Departamento de Estatística - UFMG*, 2008.

BARTLETT, M. Statistical estimation of density functions. *Sankhya Ser.*, v. 25, p. 245–254, 1963.

BERTRAND-RETALI, R. Convergence uniforme d'une estimateur de la densité par la method du noyau. *Revue Roumaine de Mathematiques Pures et Appliquées*, v. 23, n. 1, p. 361–385, 1978.

BESSEGATO, L. F. *Escolha do Parâmetro de Suavidade na Estimativa da Função de Distribuição*. Belo Horizonte, MG, Brasil: [s.n.], 2001.

BESSEGATO, L. F.; ATUNCAR, G. S.; DUCZMAL, L. H. Rotinas em r para técnicas de suavização por núcleos estimadores. *Departamento de Estatística - UFMG*, 2001.

BOWMAN, A.; AZZALINI, A. *Applied Smoothing Techniques for Data Analysis*. 1th. ed. Oxford: Oxford University Press, 1997.

BOWMAN, A. W. An alternative method of cross-validation for smoothing of density estimates. *Biometrika*, v. 71, p. 353–360, 1984.

BOWMAN, A. W.; HALL, P.; PRVAN, T. Bandwidth selection for the smoothing of distribution functions. *Biometrika*, v. 85, p. 799–808, 1998.

CHIU, S. T. Bandwidth selection for kernel density estimation. *Annals of Statistics*, v. 19, p. 1883–1905, 1991.

DAMASCENO, E. C. Projeto de Diplomação, *Escolha do parâmetro de suavidade em estimação funcional*. Belo Horizonte, MG, Brasil: [s.n.], 2000.

DEVROYE, L.; GYORFI, L. Nonparametric density estimation: the l1 view. *New York: John Wiley*, 1978.

DUIN, R. On the choice of smoothing parameter for parzen estimators of probability density functions. *IEEE Transactions on Computing*, 1976.

EPANNECHNIKOV, V. Non-parametric estimation of a multivariate probability density. In: *Theory of Probability and Its Applications*. [S.l.]: V.14, 1969. p. 153–158.

FIX, J.; HODGES, J. Nonparametric discrimination: consistency properties. In: *Report Number 4*. [S.l.]: IUSAF School of Aviation Medicine, Randolph Field, 1951. p. 249–257.

HABBEMA, J.; HERMANS, J.; BROEK, K. A stepwise discriminant analysis program using density estimation. In: *In COMPSTAT*. [S.l.]: G. Bruckmann, Pyshica-Verlag, Viena, 1974. p. 101–110.

JOHNSON, N.; KOTZ, S. *Encyclopedia of Statistical Science*. 1th. ed. New York: John Wiley & Sons, 1988.

JONES, M.; KAPPENMAN, R. On a class of kernel density estimate bandwidth selectors. *Scandinavian Journal of Statistics*, v. 19, n. 4, p. 337350, 1992.

LUCAMBIO, F. Estimador kernel da função de densidade. *Universidade Federal do Paraná*, v. 1, p. 1–10, 2008.

NADARAYA, E. On the integral mean square error of some nonparametric estimates for the density function. *Theory of Probability and It's Applications*, v. 19, p. 133–141, 1974.

PARZEN, E. On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, v. 33, p. 1065–1076, 1962.

R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2014. Disponível em: <http://www.R-project.org/>.

ROSENBLAT, M. Remarks on some nonparametric estimate of a density function. *Annals of Mathematical Statistics*, v. 1, p. 832–837, 1956.

RUDEMO, M. Empirical choice of histograms and kernel density estimators. *Journal of Statistical*, v. 9, p. 65–78, 1982.

SCHUSTER, E.; GREGORY, G. On the nonconsistency of maximum likelihood density estimators. In: *Proceeddings of the Thirteenth Interface of Computer Science and Statistcs, W. F. Eddy*. [S.l.]: New York: Springer-Verlag, 1981. p. 292–294.

SILVERMAN, B. W. *Density Estimation for Statistics and Data Analysis*. 1th. ed. New York: Chapman & Hall, 1986.

TERRELL, G. R.; SCOTT, D. W. Oversmoothed nonparametric density estimates. *J. Amer. Statist. Assoc.*, v. 80, p. 209–214, 1985.

WAND, M. P.; JONES, M. C. *Kernel Smoothing*. Chapman & Hall: New York, 1994.

WOODROOFE, M. On choosing a delta sequence. *Annals of Mathematical Statistics*, v. 41, p. 1665–1771, 1970.

5 Apêndice

```
-----------------------------------------------------
ALGORITMO-1  (Densidade - Transformação)
-----------------------------------------------------

########ENTRADA DAS VARIÁVEIS#################
amos=c(15,20,30,50,100)
for (w in 1:length(amos)){

media = 5#trocar média;
dp = 0.25#trocar desvio padrão
n = amos[w] #trocar
x= rnorm(n,media,dp)
dp.a = sqrt(var(x))
media.a = mean(x)

##############################################

l.sup = 4*dp.a+media.a; l.inf = -4*dp.a+media.a

################################################
####  TRANSFORMAÇÃO QUANDO A VARIÂNCIA NÃO É GRANDE OU PEQUENA: ###
#### SE O DESVIO PADRÃO ESTÁ ENTRE 0,3 E 3 FAÇA :##################
#################################################################
if ( dp.a>=0.3 & dp.a<3){

########VARIÁVEIS AUXILIARES ###################################
```

```
janela = plug.smooth(x)$tab$Janela ###### BUSCA A JANELA DO PLUG IN LUPERCIO
x.s = sort(x) # ordenando a variável y
int = (8*dp.a)/(n) # tamanho do intervalo da abcissa
abc = seq(l.inf,l.sup,int) # gerando os intervalos na abcissa
x.s[length(x.s)+1] = x.s[length(x.s)] # igualando ao tamanho de ABC
soma.fhat = c(0) # variávelpara a soma
fhat = c(0) # F estimado
vc.soma =c(0); vc = c(0)

#######cÁLCULO DO Fn ESTIMADO, USANDO COMO NÚCLEO k A NORMAL (0,1)#

for(i in 1:length(abc)){
for(j in 1:length(abc)){
soma.fhat[j] = dnorm((abc[i]-x.s[j])/janela)

if( i!=j){
vc.soma[j] = dnorm((abc[i]-x.s[j])/janela)
}
}

fhat[i] = sum(soma.fhat)/(length(x.s)*janela)  #cáculo do núcleo estimador

}
######################### FIM DO CÁLCULO DO FN ##################

#####################################################################
##### FIM PARA DESVIO PADRÃO ENTRE 0,3 E 3 ###################
#####################################################################
#### SE O DESVIO PADRÃO ESTÁ ENTRE 0,3 E 3 FAÇA :################
#####################################################################
if (dp.a >=3 | (dp.a<0.3 & dp.a>=0)){

####################### VARIÁVEIS AUXILIARES #################

x.i = x
```

```
tam = 100 # tamanho da malha

################################
#interpolação Aritmética#######
################################

r = (l.sup-l.inf)/(tam-1)

w.i = c(0);w.i[1]=l.inf

for(i in 2:tam){
w.i[i]= w.i[i-1]+r
}
################################

v.i = w.i/dp.a; y.i = x.i/dp.a

janela = plug.smooth(y.i)$tab$Janela ###### BUSCA A JANELA DO PLUG IN LUPERCIO
soma.fhat = c(0) # variávelpara a soma
fhat.t = c(0) # F estimado

####### cÁLCULO DO Fn ESTIMADO, USANDO COMO NÚCLEO k A NORMAL (0,1)###

for(i in 1:length(v.i)){
for(j in 1:length(y.i)){
soma.fhat[j] = dnorm((v.i[i]-y.i[j])/janela)



}

fhat.t[i] = sum(soma.fhat)/(length(x)*janela)  #cáculo do núcleo estimador

}
######################### FIM DO CÁLCULO DO FN ########################
```

```
fhat = (fhat.t)/dp.a;
abc  = w.i
#plot(w.i,fhat.t)

}

#########################################################################
#######FIM PARA DESVIO PADRÃO ENTRE 0,3 E 3 ##############################
#########################################################################

#################### COMANDOS GRÁFICOS: ######################

xx=plug.smooth(x)$x
#fhat = fhat[1:length(x.i)]
est=plug.smooth(x)$fx
real=dnorm(xx,mean=media,sd=sqrt(dp**2))
max2=max(est,real)

win.graph()

plot(xx, est,  main=bquote(paste("Normal(", mu, "=", .(media),"; ", sigma, "=", .(d
n=",.(n))), type="l", col="blue", lty=1, lwd=2, ylab="f(x)", xlab="x", ylim=c(0,max

lines(xx,real, col="red", lty=2, lwd=2)

grid()

legend("topright", legend = c(" Estimado L", "Real", "Fn - estimado", "Val. Cruz.")
lty = c(1,2), lwd=c(2,2), col=c("blue","red","green","black"), bty="n")

}
```

```
-----------------------------------------------------
ALGORITMO-2  (Distribuição - Transformação)
-----------------------------------------------------



amostra = c(500,100,50,15)
desvio =c(0.01,.2,3.5,10);
for (a in 1:length(amostra)){
for(b in 1:length(desvio)){

n = amostra[a] #trocar
dp =desvio[b]#trocar desvio padrão
media = 10#trocar média;

x= rnorm(n,media,dp) ;dp.a = sqrt(var(x)); media.a = mean(x)

#############################################################################

l.sup = 4*dp.a+media.a; l.inf = -4*dp.a+media.a
lambda = alg.lambda(x);
H = alg.H(x,lambda)$h
h.opt = alg.hop.F(x,H)

#############################################################################
####  TRANSFORMAÇÃO QUANDO A VARIÂNCIA NÃO É GRANDE OU PEQUENA: ##################

#### SE O DESVIO PADRÃO ESTÁ ENTRE 0,3 E 3 FAÇA :#################################
#############################################################################

if ( dp.a>=0.3 & dp.a<3){

####################### VARIÁVEIS AUXILIARES ###################################

x.s = sort(x) # ordenando a variável y
```

```
int = (8*dp.a)/(n) # tamanho do intervalo da abcissa
abc = seq(l.inf,l.sup,int) # gerando os intervalos na abcissa
x.s[length(x.s)+1] = x.s[length(x.s)] # igualando ao tamanho de ABC
soma.fhat = c(0) # variávelpara a soma
fhat = c(0) # F estimado

#### cÁLCULO DO Fn ESTIMADO, USANDO COMO NÚCLEO k A NORMAL (0,1)##################

for(i in 1:length(abc)){
for(j in 1:length(abc)){
soma.fhat[j] = pnorm((abc[i]-x.s[j])/h.opt)

}

fhat[i] = sum(soma.fhat)/(length(x.s))  #cáculo do núcleo estimador

}
########################## FIM DO CÁLCULO DO FN ###################################

}

###########################################################################################
###################### FIM PARA DESVIO PADRÃO ENTRE 0,3 E 3 ###################
###########################################################################################
#### SE O DESVIO PADRÃO ESTÁ ENTRE 0,3 E 3 FAÇA :###################################
###########################################################################################

if (dp.a >=3 | (dp.a<0.3 & dp.a>=0)){

####################### VARIÁVEIS AUXILIARES ###################################

x.i = x
tam = 100 # tamanho da malha

################################
```

```
#interpolação Aritmética#######
##############################

r = (l.sup-l.inf)/(tam-1)

w.i = c(0);w.i[1]=l.inf


for(i in 2:tam){
w.i[i]= w.i[i-1]+r
}
#####################################

v.i = w.i/dp.a; y.i = x.i/dp.a

### Encontrando a janela da variável transformada

lambda = alg.lambda(y.i);
H = alg.H(y.i,lambda)$h
h.opt = alg.hop.F(y.i,H)


soma.fhat = c(0) # variávelpara a soma
fhat.t = c(0) # F estimado

##### cÁLCULO DO Fn ESTIMADO, USANDO COMO NÚCLEO k A NORMAL (0,1)#################

for(i in 1:length(v.i)){
for(j in 1:length(y.i)){
soma.fhat[j] = pnorm((v.i[i]-y.i[j])/h.opt)


}

fhat.t[i] = sum(soma.fhat)/length(x)  #cáculo do núcleo estimador
```

```
}
######################### FIM DO CÁLCULO DO FN ###################################

fhat = (fhat.t);
abc  = w.i
#plot(w.i,fhat.t)

}
##########################################################################################
###################### FIM PARA DESVIO PADRÃO ENTRE 0,3 E 3 ###################
##########################################################################################
#################### COMANDOS GRÁFICOS: ##################################

xx=plug.smooth(x)$x
#fhat = fhat[1:length(x.i)]
#est=plug.smooth(x)$fx
real=pnorm(xx,mean=media,sd=sqrt(dp**2))
max2=max(fhat,real)

win.graph()
plot( xx, real,  main=bquote(paste("Normal(", mu, "=", .(media),"; ",
sigma, "=", .(dp),");n=", .(n))), type="l",  col="red", lty=1, lwd=2,
ylab="P(x)", xlab="x", ylim=c(0,max2))

lines(abc,fhat,type="l", col="green", lty=2, lwd=2)
grid()
legend("topleft", legend = c("Real", "Fn - estimado"), lty = c(1,2),
lwd=c(2,2),col=c("red","green"), bty="n")
}
}
```

```
-------------------------------------------------------
ALGORITMO-3  (Histograma - Janela Ótima - Distribuição)
-------------------------------------------------------


#############################################################################
n.amostra= c(30,50,100,500) # tamanhos amostrais

repet = 100 # Repetições Simulação
media = 10 # média para a simulação
desv = c(.1,0.5,2,3,3.5,5) #vetor de desvios
n.des = length(desv) # tamanho do vetor de desvios
cont = 0  #contador de transformações

h.opt=array(0,dim = c(length(n.amostra),repet)) #matriz das janelas -simulações
v.ar = array(0,dim=c(length(n.amostra),5))

for (a in 1:n.des){ dp = desv[a]#
for (b in 1:length(n.amostra)){ n=n.amostra[b]#b in 1:n.des){ dp = desv[b]
for(c in 1:repet){
x = rnorm(n,media,dp)
dp.a = sqrt(var(x))

#### SE O DESVIO PADRÃO ESTÁ ENTRE 0,3 E 3 FAÇA :#################################
#############################################################################

if ( dp.a>=0.5 & dp.a<3){

lambda = alg.lambda(x);
H = alg.H(x,lambda)$h
h.opt[b,c] = alg.hop.F(x,H)

}
```

```
#### SE O DESVIO PADRÃO ESTÁ ENTRE 0,3 E 3 FAÇA :#################################
################################################################################
if (dp.a >=3 | (dp.a<0.5 & dp.a>=0)){

######################## VARIÁVEIS AUXILIARES#########################################

x.i = x; y.i = x.i/dp.a

lambda = alg.lambda(y.i);
H = alg.H(y.i,lambda)$h
h.opt[b,c] = alg.hop.F(y.i,H)
cont = cont+1
}

} #fim da repetição (C)


#############c comandos gráficos####################

v.ar[b,1] = mean(h.opt[b,]); #média
v.ar[b,2]  = sqrt(var(h.opt[b,])) #dp
v.ar[b,3] = v.ar[b,1] - v.ar[b,2] # media - dp
v.ar[b,4] = v.ar[b,1] + v.ar[b,2] # media + dp
v.ar[b,5] = cont*100/repet # contador de transformações em %
cont = 0

} # fim do n.amostra.(B)

x.min = min(v.ar[,3])-v.ar[b,2] ; x.max = max(v.ar[,4])+v.ar[b,2]

for(k in 1:length(n.amostra)){

win.graph()
hist((h.opt[k,]), prob=TRUE,xlab ="h.ótima",ylab = "Probabilidade",
main=bquote(paste("Normal(", mu, "=", .(media),"; ", sigma, "=",
```

```
.(desv[a]),"); n=", .(n.amostra[k]), "; Rep. =", .(repet), ";
Transf = ", .(v.ar[k,5]),"%")),xlim=c(x.min,x.max))

abline(v=v.ar[k,1], col = "red", lty=1, lwd=2); abline(v=v.ar[k,3],
col = "green", lty=2, lwd=1);abline(v=v.ar[k,4], col = "green",
lty=2, lwd=1)



}
h.opt=array(0,dim = c(length(n.amostra),repet))
v.ar = array(0,dim=c(length(n.amostra),5)); cont = 0
}  # fim do (desv)(A)
```

```
-----------------------------------------------------
ALGORITMO-4  (Histograma - Janela Ótima - Distribuição)
-----------------------------------------------------


################################
dp.a = sqrt(var(x.i)); media  = mean(x.i)
y.i = x.i/dp.a

janela1 = plug.smooth(y.i)$tab$Janela

l.inf =  min(x.i)-janela1
l.sup = max(x.i)+ janela1#limite superior e inferior
################################
################################
#interpolação Aritmética#######
################################

r = (l.sup-l.inf)/(tam-1)
w.i = c(0);w.i[1]=l.inf

for(i in 2:tam){
w.i[i]= w.i[i-1]+r
}
#####################################
v.i = w.i/dp.a

janela = plug.smooth(v.i)$tab$Janela ####
## BUSCA A JANELA DO PLUG IN LUPERCIO
soma.fhat = c(0) # variávelpara a soma
fhat = c(0) # F estimado
```