Universidade Federal de Minas Gerais

Instituto de Ciências Exatas

Departamento de Estatística

Núcleo Estimador para Função de Densidade e Função Quantílica na Presença de Censura à Direita

Orientando: Carlito A. S. Balbino

Orientador: Prof. Gregorio Saraiva Atuncar

março 2006

Agradecimentos

A Deus, pelo o dom da vida e força para vencer os momentos difíceis.

Aos meus pais e a minha irmã pela educação e formação, e por toda paciência e apoio que sempre me deram.

Aos amigos que fiz nos tempos de graduação, Mercio, Ronaldo, Ricardo, pela amizade e incentivo.

Aos professores da graduação em Matemática pela minha formação, em especial aos professores Olímpio Hiroshi Miyagaky e Margareth da Silva Alves pelo incentivo constante.

Ao professor Gregório, pela oportunidade, confiança e orientação que tornaram possível a realização deste trabalho.

Ao Flávio, Jaqueline e Eduardo, pelas amizades e oportunidades de emprego, que me ajudaram a chegar ao fim deste trabalho.

A todos os professores e funcionários do departamento, que se fizeram sempre presentes.

A todos os colegas do mestrado que participaram junto comigo desta caminhada, compartilhando os bons e maus momentos, em especial a Taynana, Josenete, Ricardo, Alexandre, Janaína, Juliana, Polyana, Marcos e Cristiano, pelo convívio dia a dia no Icex.

Aos amigos, Max e Geraldo pela preocupação e paciência no longo tempo de república, Fábio e Luciano pelo convívio tranquilo no dia a dia da república, em geral pela a oportunidade de ter conhecido pessoas verdadeiras e sempre presentes.

A todos aqueles que, direta ou indiretamente, colaboraram para a realização deste trabalho.

Sumário

Lista de Figuras

Lista de Tabelas

Resumo

O estudo do método do núcleo é bastante difundido na estimação não-paramétrica para várias funções na estatística. Dentre elas, tem-se a função de densidade e a função quantílica, discutidas nesta dissertação para dados censurados à direita. A questão principal neste método é a escolha do parâmetro de suavização h, conhecido na literatura como janela. O método de validação cruzada para seleção de h tem sido bastante estudado na estimação da função de densidade. Nesta dissertação, encontra-se uma avaliação do método usando distribuição weibull. A janela selecionada por este método subestima a janela teórica dada em Marron e Padgett[1987]. Em Chagas[2004] foi feito o mesmo estudo com a distribuição normal chegando as mesmas conclusões.

Para a função quantílica, o método do Bootstrap é empregado na seleção de h. A janela selecionda é dada pelo menor erro quadrático médio através das amostras bootstrap. São avaliados neste texto alguns intervalos de confiança propostos em Padgett e Thombs[1986] e Cheng[2002]. Nas simulações realizadas para as distribuições weibull e lognormal para vários tamanhos de amostras, percebeu-se que para percentis menores ou iguais a mediana, o intervalo de confiança dado em Cheng[2002] sobressaiu aos intervalos de confiança dados por Padgett e Thombs[1986]. Já para o percentis maiores que a mediana não houve diferença entre os intervalos de confiança.

Encontra-se ainda no final desta dissertação, o algortimo para o cálculo dos estimadores.

Palavras Chaves: *Núcleo Estimador, Método Bootstrap, Método da Validação Cruzada, Função de Densidade, Função Quantílica, Intervalo de Confiança.*

Abstract

The kernel method is largely used in nonparametric estimation of many statistical functions. Among them are the density function and the quantile function discussed in this dissertation for right-censored data. The main issue in this method is the choice of the bandwidth h. The cross-validation method for the selection of h has been largely studied in the estimation of the density function. This dissertation presents an evaluation of the method in the case of the Weibull distribution. The bandwidth selected by this method underestimates the theoretical one given in Marron and Padgett (1987). Chagas (2004) conducted the same study with the normal distribution reaching the same conclusions.

For the quantile function, the Bootsrap method is used in the selection of h. The selected bandwidth is given by the smallest mean quadratic error found in the bootstrap samples. Some confidence intervals proposed in Padgett and Thombs (1986) and Cheng (2002) are evaluated in this context. In the simulations run for the Weibull and lognormal stributions for various sample sizes, it was observed that for quantiles smaller or equal to the median, the confidence interval given in Cheng (2002) were better than the confidence intervals given by Padgett and Thombs (1986). However, for quantiles greater than the median no difference was observed between those methods.

The algorithm for the calculation of the estimators can be found at the end of this dissertation.

Keywords: Kernel Estimator, Bootstrap Method, Cross-validation Method, Density Function, Quantile Function, Confidence Interval.

Capítulo 1

Introdução e Preliminares

1.1 Introdução

Tanto em estudos médicos quanto na área industrial, é comum o surgimento de dados censurados. A busca de estimadores não-paramétricos da função de densidade e da função quantílica para estes dados é de suma importância.

Devido às propriedades assintóticas, o método do núcleo tem sido bastante usado. O grande problema apresentado por este método é referente à escolha do parâmetro de suavização h, que chamaremos de janela. Torna-se, então, interessante estudar estimadores da janela ótima para a obtenção da função que exprima a probabilidade de ocorrência dos dados.

No caso da função de densidade, foi usado o método da Validação Cruzada por Mínimos Quadrados, adaptado para dados censurados, para selecionar a janela. A janela h selecionada é resultado do menor erro quadrático integrado. A partir de resultados de simulações, observou-se que tal método apresenta variabilidade muito grande, isto compromete o desempenho na estimação da densidade.

Na estimação da função quantílica, o método do Bootstrap foi empregado para seleção da janela. Esta janela selecionada é resultado do estudo comparativo feito por Padgett[1986]. Escolhia-se a janela através da maior razão entre o erro quadrático médio do estimador de Kaplan-Meier quantílico e o erro quadrático médio da função quantílica estimado através do núcleo.

Aproximações para intervalos de confiança foram calculados também.

Trabalhou-se com três tipos de intervalos, dois propostos por Padgett e Thombs[1986] e o outro por Cheng[2002]. Todos estes três intervalos usam a janela selecionada através do método do Bootstrap para a função quantílica.

Em suma, este trabalho propõe-se a avaliar a janela selecionada para a função de densidade e a função quantílica e comparar três métodos para aproximação de intervalos de confiança para os quantis.

1.2 Revisão da Literatura

A busca por bons estimadores, sejam eles paramétricos ou não paramétricos, tem sido objeto de estudo ao longo do tempo. Para dados completos, existe uma ampla literatura de estimadores não paramétricos para a função de densidade através do método do núcleo, introduzido por Rosenblatt[1956]. Parzen[1962] demonstrou algumas propriedades assintóticas destes estimadores. Mais resultados podem ser encontrados em Silverman[1986] e Simonoff[1996].

No caso de dados censurados à direita, a busca de estimadores para as funções de densidade e quantílica através do núcleo foi objeto de estudo de Blum e Sursala[1980], Padgett e McNichols[1984], Marron e Padgett[1987], Padgett[1986], Padgett e Thombs[1986], Cheng[2002].

Blum e Sursala[1980] apresentaram, para dados censurados, um estimador alternativo para função de densidade através do núcleo originado dos resultados de Rosenblatt[1976], que será visto na seção 2.2.3.

Para medir a eficiência global das estimativas, é comum usar o *erro quadrático médio integrado*($EQMI$). Uma outra medida também usada é o *erro quadrático integrado*(EQI).

Três vantagens são apresentadas por Marron e Padgett[1987] em se trabalhar com o EQI ao invés do $EQMI$ para o caso dos dados censurados à direita.

1. O $EQMI$ não existe para o estimador baseado na função produto limite;

2. O EQI é um critério de erro mais conservador, porque avalia a função densidade para a amostra fixa, em vez de só avaliar em cima de todos os dados possíveis;

3. O EQI é mais simples para a seleção de janela usando o método da Validação Cruzada.

Na presente dissertação, $\hat{f}_n(x)$ denota o estimador da função de densidade usando o método do núcleo. Esse estimador envolve na sua definição um núcleo K a ser escolhido entre alguns a serem citados posteriomente e o parâmetro de suavidade h, conhecido na literatura como janela, motivo deste estudo. Em geral, h é escolhido de maneira que $\hat{f}_n(x)$ seja um estimador ótimo de f^0, de acordo com alguma medida de desempenho, onde f^0 é a função de densidade.

Marron e Padgett [1987] mostraram que $\sup_h \left| \frac{EQI(f_n) - EQMI(f_n)}{EQMI(f_n)} \right| \xrightarrow{qc} 0$. Também, sob algumas suposições que serão vistas posteriomente, apresentaram uma janela ótima para o estimador da função de densidade,

$$h_o = \left[\frac{(\int K^2) \int (f^o/H)}{2\mathsf{k}(\frac{\kappa}{\mathsf{k}!})^2 [\int ((f^o)^{(\mathsf{k})})^2]} \right]^{\frac{1}{2\mathsf{k}+1}} n^{-\frac{1}{2\mathsf{k}+1}} \tag{1.1}$$

com taxa de convergência $EQI \approx n^{-\frac{2\mathsf{k}}{2\mathsf{k}+1}}$, onde κ é o k-ésimo momento da função núcleo K e f^o é a função de densidade teórica(mais detalhes no teorema 2.2.1).

O grande problema para (1.1) é que a janela depende da função de densidade desconhecida f^o.

Para contornar este problema, Marron e Padgett[1987] adaptaram, para dados censurados, o método de Validação Cruzada por Mínimos Quadrados introduzido por Rudemo [1982] e Bowman [1984] para dados completos.

Definindo $\hat{h}_c$ como sendo o minimizador pelo método da Validação Cruzada por mínimos quadrados, Padgett e Marron[1987] também mostraram que a razão

$$\frac{EQI(\hat{f}_n, \hat{h}_c)}{\inf_h EQI(\hat{f}_n, h)} \xrightarrow{qc} 1.$$

Assim, assintoticamente, o h escolhido através da Validação Cruzada por Mínimos Quadrados é um estimador ótimo.

Usando amostras da distribuição normal, Chagas[2004] calculou a janela ótima em (1.1) proposta por Marron e Padgett[1987] e comparou-a com a janela encontrada atráves do método da Validação Cruzada, observando que:

- O estimador da janela ótima obtido pelo método de Validação Cruzada para dados censurados é assintoticamente normal, mas apresenta grande variabilidade;

- Mantendo a proporção de censura constante e aumentado a janela, a suavidade da curva de densidade aumenta. Porém, nem sempre a curva mais suavizada foi a melhor estimativa da função de densidade;

- O método torna-se impreciso à medida que aumenta a proporção de dados censurados.

Em Biao Zhang [1996], encontra-se mais resultados assintóticos para estimação da densidade através do núcleo.

Uma outra função de interesse é a função quantílica, denotada por $Q^0(p)$, para $p \in (0,1)$. Sander[1975] propôs a estimação da função quantílica pelo estimador produto limite, definido por $\tilde{Q}_n = \tilde{F}_n^{-1}$, onde $\tilde{F}_n$ é o estimador produto limite da função distribuição do tempo de vida F^0 (Kaplan-Meier[1958] e Efron[1967]). Sander[1975] e Cheng[1984] conseguiram algumas propriedades assintóticas para $\tilde{Q}_n$ e Csörgo[1983] discutiu fortes aproximações dos resultados. A função quantílica do estimador não-paramétrico produto limite é uma função escada com saltos correspondentes às observações não-censuradas.

Na procura de um estimador mais suave, um estimador não-paramétrico para a função quantílica para dados censurados à direita baseado na função núcleo foi proposto em Padgett[1986], originado dos resultados para amostras completas de Yang[1985]. Lio, Padgett e Yu[1986] estudaram algumas propriedades assintóticas deste estimador.

A técnica do bootstrap foi usada por Padgett e Thombs[1986] para selecionar a janela h.

O bootstrap também é usado para estimar intervalos de confiança usando a janela já selecionada na estimativa pontual. Em Padgett e Thombs[1986], a proposta de estimar intervalos de confiança é através do método núcleo usando a janela já selecionada na estimativa pontual, onde foi usada a estimativa da variância bootstrap na seleção da janela para definir uma aproximação para o intervalo de confiança para $Q^0(p)$,

$$\left[Q_n(p) - z_{(1-\frac{\alpha}{2})}\sqrt{\hat{V}ar(h)},\ Q_n(p) + z_{(1-\frac{\alpha}{2})}\sqrt{\hat{V}ar(h)}\right]$$

onde $z_{1-\frac{\alpha}{2}}$ é o $(1-\frac{\alpha}{2})$ percentil da distribuição normal padrão e $\hat{V}ar(h)$ é variância das estimativas dos quantis $\hat{Q}_{n,i}(p)$ através do bootstrap, com $i = 1, \ldots, B$ onde B é

o número de amostras bootstrap.

Efron[1986] demonstrou que o intervalo simétrico da forma $\left[\hat{\theta} \pm z\sigma\right]$ é correto se a estatística $\hat{\theta}$ apresenta distribuição normal.

Lio, Padgett e Yu[1986] demonstraram a normalidade assintótica para $Q_n(p)$. Padgett e Thombs[1986] observaram que para amostras pequenas e até moderadamente grandes e percentil próximo de 0 ou 1, a distribuição de $Q_n(p)$ apresentava-se de forma assimétrica, tornando o intervalo impreciso.

Uma outra aproximação intervalar sugerida por Padgett e Thombs[1986] é usar a distribuição empírica das estimativas bootstrap de $Q_n(p)$.

Em Cheng[2002], um terceiro intervalo é proposto onde usa-se a derivada da função quantílica na aproximação do intervalo. Resultados para a derivada da função quantílica podem ser encontrados também em Cheng[2002].

Uma nova proposta é encontrada em Cao e Jácome[2003], onde a variável indicadora de censura δ é trocada pelo estimador não paramétrico da probabilidade condicional $p(t) = E(\delta|T = t)$.

1.3 Definições Básicas em Análise de Sobrevivência

A análise de sobrevivência é normalmente definida como o conjunto de técnicas e modelos estatísticos usados na análise do comportamento de uma variável aleatória T positiva na presença de **censura** (observação parcial da variável resposta). Na maioria das vezes, a variável resposta é o tempo até a ocorrência de um evento de interesse (denominado "tempo de falha") que pode ser, por exemplo, o tempo até a morte de um paciente ou até a cura ou reciditiva(reincidência) de uma doença.

Censura é a característica em estudos de dados em sobrevivência. Mesmo censurados, todos os resultados obtidos através de um estudo de sobrevivência devem ser incorporados na análise estatística. A justificativa para tal procedimento se baseia nas seguintes questões:

- Mesmo sendo incompletas, as observações censuradas nos fornecem alguma informação a respeito da variável resposta;

- A omissão das censuras no cálculo das estatísticas de interesse certamente levará

a conclusões viciadas(Colosimo[2001]).

Existem três formas de censura. São elas:

- **Censura à direita** : caracteriza-se pelo fato de que o tempo de ocorrência do evento de interesse é maior do que tempo registrado;
- **Censura à esquerda** : ocorre quando o tempo registrado é maior que o tempo de falha, ou seja, o evento de interesse já aconteceu quando o indivíduo foi observado;
- **Censura intervalar** : caracteriza-se pelo fato de que em alguns conjuntos de dados, o valor exato da variável resposta não é observado, sendo sabido apenas que seu valor pertence a um certo intervalo.

O tipo de censura usado neste trabalho é a censura à direita oriunda de uma variável aleatória independente do tempo de vida. Censura não informativa.

Definição 1.3.1. *A função de Sobrevivência $S(t)$ é definida como a probabilidade de uma observação não falhar até o tempo t, ou seja:*

$$S(t) = P(T > t) = 1 - F^0(t),$$

onde F^0 é a função de distribuição de T.

Uma característica da distribuição de sobrevivência importante é a função quantílica que é útil em confiabilidade e estudos médicos.

Definição 1.3.2. *A função de Quantílica $Q^0(p)$ é definida como*

$$Q^0(p) = (F^0)^{-1}(p) = \inf\left\{t : F^0(t) \geq p\right\}, \ \ 0 \leq p \leq 1,$$

onde $F^0(t)$ é a função distribuição de T.

1.4 Métodos Bootstrap

Para medir a precisão dos estimadores, é necessário calcular algumas medidas de variabilidade, o que na maioria das vezes é feito utilizando o erro padrão dos mesmos. Em algumas situações (por exemplo, situações onde não conhecemos a

distribuição exata do estimador), o cálculo dessas medidas torna-se muito complicado. Nesse contexto, o método do Bootstrap, criado por Efron[1979], surge como uma solução para esse tipo de problema, uma vez que ele é um método computacionalmente intensivo, e portanto, pode ser utilizado independentemente de dificuldades teóricas ou matemáticas envolvendo o estimador do parâmetro em questão.

Esse método consiste em uma técnica de reamostragem que permite aproximar a distribuição de uma função das observações pela distribuição empírica dos dados, com base em uma amostra de tamanho finito. A amostragem é feita com reposição da amostra original (bootstrap não-paramétrico). Nessa situação, supomos que as observações são obtidas da função de distribuição empírica, $\hat{F}$, atribuindo uma probabilidade igual a $1/n$ para cada ponto amostral.

Atualmente, devido aos recursos computacionais disponíveis, o método do Bootstrap, além de ser usado para medir a precisão de estimadores, é aplicado também na construção de intervalos de confiança, testes de hipóteses, estimação do vício, etc. Vamos definir a amostra bootstrap seguindo a notação de Efron e Tibshirani [1993].

Definição 1.4.1. (Bootstrap não paramétrico) *Sejam $X_1, X_2, \ldots, X_n$ variáveis aleatórias independentes com função de distribuição comum F. Se F é desconhecida, assumimos que a função geradora dos dados seja tal que, a cada um dos n valores da amostra seja atribuída uma probabilidade $1/n$ para sua ocorrência. Deste modo, a amostra bootstrap é construída retirando-se, com reposição, um conjunto de n observações da amostra original, que igualmente denotamos por $X^\star = X_1^\star, \ldots, X_n^\star$.*

Suponha agora que nossa estatística de interesse é dada por $T(X_1, \ldots, X_n)$. O procedimento bootstrap básico consiste em gerar repetidas amostras bootstrap, cada uma delas de tamanho n, um número suficientemente grande de vezes. Obtemos assim, as amostras bootstrap $X^{\star 1}, \ldots, X^{\star B}$, onde B é o número de replicações bootstrap.

O valor ideal para B depende da finalidade para qual a teoria está sendo utilizada. Por exemplo, se desejamos somente estimar o erro padrão , Efron e Tibshirani [1993] sugerem valores de B entre 50 a 200 se o interesse é construir intervalos de confiança, Hall [1986] faz um estudo do número de replicações necessárias para se obter os níveis

de significância desejados e sugere tomar a probabilidade de corbertura nominal como múltiplo de $(B+1)^{-1}$.

Para cada amostra bootstrap, calculamos o valor da estatística de interesse a qual denotaremos por $T^\star(X^{\star\ b}),\ b = 1, \ldots, B$. Assim, podemos usar a distribuição empírica de $T^\star(X^\star)$ como uma aproximação para a verdadeira distribuição da estatística $T(X)$, e portanto fazer inferências sobre o parâmetro θ. O procedimento acima é referido como princípio *"plug-in"* por Efron e Tibshirani [1993].

1.4.1 A Estimativa Bootstrap do Erro Padrão

Como vimos anteriormente, o método do Bootstrap é um método que tem por finalidade principalmente determinar a precisão de estimadores para os quais uma forma analítica de erro padrão é muito complicada ou até mesmo inexistente.

Por exemplo, vamos supor que estamos interessados em estimar o erro padrão de uma estatística de interesse. O primeiro passo é construir as amostras bootstrap através de amostragem com reposição da amostra original. Assim, se dispomos de n observações $X_1, X_2, \ldots, X_n$, a primeira amostra bootstrap, que denotamos por $X^{\star 1}$, é obtida retirando-se, com reposição, n elementos desse conjunto de dados. Para essa amostra, calcula-se o valor da estatística de interesse T, que nesse caso será denotada por $T^\star(X^{\star 1})$. Repeti-se esse procedimento um número grande, B, de vezes e, para cada amostra bootstrap, calculamos $T^\star(X^{\star b}),\ b = 1, \ldots, B$. A estimativa bootstrap do erro padrão de $T(X)$ é então definida como o desvio padrão amostral das B replicações bootstrap, e é dada por :

$$\hat{ep}(T^\star) = \left[\frac{\sum_{b=1}^{B} [T^\star(X^{\star\ b}) - T^\star(.)]^2}{B-1} \right]^{1/2}$$

em que $T^\star(.) = \dfrac{\sum_{b=1}^{B} T^\star(X^{\star\ b})}{B}$.

Capítulo 2

Núcleo Estimadores

2.1 Generalidades no Caso de Dados Completos

2.1.1 Introdução

O método do núcleo é um método não paramétrico bastante difundido para dados completos na estimação de funções densidade de probabilidade. O método foi introduzido na literatura por Rosenblatt[1956] e Parzen[1962].

Definição 2.1.1. *Seja $X_1, \ldots, X_n$ uma amostra aleatória de uma distribuição com função densidade de probabilidade f. O núcleo estimador de f avaliado no ponto x é definido por:*

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right)$$

onde h é denominado parâmetro de suavização ou janela (bandwidth) e K é uma função denominada núcleo.

Geralmente, a escolha do parâmetro de suavização h é muito mais importante para um melhor desempenho desse método do que a escolha da função núcleo K (Silverman[1986]).

Alguns núcleos conhecidos são apresentados a seguir, sendo os dois primeiros usados na dissertação.

Triangular	$K(x) = (1 - \|x\|), \quad -1 \leq x \leq 1$
Gaussiano	$K(x) = \frac{1}{\sqrt{2\pi}} exp\left(-\frac{x^2}{2}\right), \quad -\infty < x < \infty$
Uniforme	$K(x) = 1, \quad -\frac{1}{2} \leq x \leq \frac{1}{2}$
Epanechnikov	$K(x) = 0,75(1 - x^2), \quad -1 \leq x \leq 1$

2.1.2 Medidas de Desempenho do Núcleo

Nesta seção serão estudadas algumas medidas para avaliar a qualidade do núcleo estimador, que seram usadas como critério para encontrar o h ótimo denotado por "h_{opt}".

Considerando um ponto x fixo, uma medida do desempenho de $\hat{f}$ é o *erro quadrático médio*(EQM) definido por

$$EQM[\hat{f}(x)] = E\left[\left(\hat{f}(x) - f(x)\right)^2\right]. \tag{2.1}$$

Uma observação interessante é que a expressão (2.1) pode ser escrita em função do vício e da variância de $\hat{f}(x)$:

$$EQM[\hat{f}(x)] = Var\left[\hat{f}(x)\right] + \left\{E\left[\hat{f}(x)\right] - f(x)\right\}^2.$$

O problema para o EQM é que esta medida é de avaliação pontual. Se for desejado uma medida global, precisa-se de um critério de erro que mede globalmente a distância entre as funções $\hat{f}$ e f. Tal critério de erro é o *erro quadrático integrado*(EQI) dado por:

$$EQI[\hat{f}(x)] = \int \{\hat{f}(x) - f(x)\}^2 dx.$$

O EQI é apropriado somente se a preocupação for com a amostra obtida. Se quisermos uma análise sobre as possíveis amostras, será mais apropriado analisar o valor esperado desta quantidade aleatória, o *erro quadrático médio integrado*$(EQMI)$ cuja definição é dada por :

$$EQMI[\hat{f}(x)] = E\int \{\hat{f}(x) - f(x)\}^2 dx. \tag{2.2}$$

Considerando que o integrando em (2.2) é não-negativo, pode-se inverter a ordem da integração com a esperança e então encontrar uma forma alternativa para o $EQMI$

dada por:

$$\begin{aligned} EQMI[\hat{f}(x)] &= E\int\{\hat{f}(x)-f(x)\}^2dx = \int E\{\hat{f}(x)-f(x)\}^2dx = \int EQM\{\hat{f}(x)\} \\ &= \int\left\{E\left[\hat{f}(x)\right]-f(x)\right\}^2dx + \int Var\left[\hat{f}(x)\right]dx. \end{aligned}$$

Examina-se agora o vício e a variância do estimador. Silverman [1986] mostra que:

$$E[\hat{f}(x)] = \int \frac{1}{h}K\left(\frac{x-y}{h}\right)f(y)dy \tag{2.3}$$

e

$$Var\left[\hat{f}(x)\right] = \frac{1}{n}\left\{\int \frac{1}{h^2}K\left(\frac{x-y}{h}\right)^2 f(y)dy - \left[\frac{1}{h}\int K\left(\frac{x-y}{h}\right)f(y)dy\right]^2\right\}. \tag{2.4}$$

Sabe-se que, em situações gerais, o vício e a variância não podem ser facilmente calculados. Porém, sob certas suposições, pode-se encontrar expressões aproximadas para (2.3) e (2.4). Dessa forma, suponha que:

1. A função núcleo K é simétrica;
2. $\int K(t)dt = 1$;
3. $\int tK(t)dt = 0$;
4. $\int t^2K(t)dt \neq 0$;
5. A função desconhecida f tem derivadas contínuas de todas as ordens necessárias.

Sabe-se que

$$Vício[\hat{f}(x)] = \frac{1}{h}\int K\left(\frac{x-y}{h}\right)f(y)dy - f(x).$$

Portanto, fazendo a mudança de variável $y = x - ht$ e utilizando a suposição (1), tem-se:

$$Vício[\hat{f}(x)] = \int K(t)f(x-ht)dt - f(x) = \int K(t)\{f(x-ht)-f(x)\}dt.$$

Por outro lado, usando expansão em série de Taylor temos:

$$f(x-ht) = f(x) - htf'(x) + \frac{1}{2}h^2t^2f''(x) + \ldots$$

e então, utilizando novamente as suposições acima, tem-se que:

$$Vício[\hat{f}(x)] \approx \frac{1}{2}h^2 f''(x)k_2. \tag{2.5}$$

onde $k_2 = \int t^2 K(t)dt$.

Quanto à variância, usando a mesma substituição $y = x - ht$ e utilizando (2.3) e (2.4), conclui-se que:

$$Var\left[\hat{f}(x)\right] = \frac{1}{nh}\int f(x-ht)K(t)^2dt - \frac{1}{n}\{f(x) + O(h^2)\}^2$$

Finalmente, utilizando expansão em série de Taylor, chega-se à seguinte expressão:

$$Var\left[\hat{f}(x)\right] \approx \frac{f(x)\int K(t)^2dt}{nh}. \tag{2.6}$$

Portanto, através do EQMI e das aproximações (2.5) e (2.6), pode-se mostrar que o valor de h que torna o EQMI mínimo é dado por:

$$h_{opt} = {k_2}^{-2/5}\left\{\int K(t)^2dt\right\}^{1/5}\left\{\int f''(x)^2dx\right\}^{-1/5} n^{-1/5}.$$

O estudo feito nesta seção foi para a função densidade para amostras completas.

2.2 Núcleo Estimadores de Funções na Presença de Censura

2.2.1 Mais Definições Básicas

Definição 2.2.1. *Um conjunto $k \subset \mathbb{R}$ é dito compacto se ele é limitado e fechado.*

Definição 2.2.2. *Uma função $f : X \to \mathbb{R}$ diz-se uniformemente contínua quando, para cada $\epsilon > 0$, existe $\delta > 0$ tal que para $x, y \in X$,*

$$|x - y| < \delta \Rightarrow |f(x) - f(y)| < \epsilon.$$

Definição 2.2.3. *Uma função $f : X \to \mathbb{R}$ é dita Hölder contínua se existem constantes $c > 0$ e $\alpha > 0$ tais que, para $x, y \in X$,*

$$|f(x) - f(y)| \leq c^{\alpha}|x - y|.$$

Definição 2.2.4. *Uma função $f : X \to \mathbb{R}$ é dita lipschitziana se existe uma constante $c > 0$ tal que, para $x, y \in X$,*

$$|f(x) - f(y)| \leq c|x - y|.$$

2.2.2 Introdução

Sejam $X_1^o, X_2^o, ..., X_n^o$ (denotando o tempo de vida) uma amostra aleatória da variável X não-negativa com função de distribuição F^o absolutamente contínua, com função de densidade f^o e função quantílica $Q^o(p)$. Independente dos $X_i^{o\prime}s$, sejam $U_1, U_2, ..., U_n$ (denotando tempo de censura) uma amostra aleatória da variável aleatória U, não-negativa com função distribuição H contínua. Denote $H^* = 1 - H$.

Sobre o modelo censurado à direita, tem-se somente o mínimo entre o X_i^o e U_i e o indicador de qual variável é a menor.

$$X_i = min\{X_i^o, U_i's\} \;\; e \;\; \delta_i = 1_{[X_i^o \leq U_i]},$$

com $1_{[.]}$ denotando a variável aleatória indicadora de ocorrência do evento [.].

Ordene os $X_i's$ acompanhados dos seus respectivos $\delta_i's$, representando-os por (Z_i, Λ_i), $i = 1, 2, ..., n$.

Na busca do estimador através da função núcleo, usa-se o conhecido estimador produto-limite**(PL)** da função sobrevivência $1 - F^o(t)$, proposto por Kaplan-Meier [1958] e demonstrado ser consistente por Efron [1967], definido por:

$$\hat{P}_n(t) = \begin{cases} 1, & 0 \leq t \leq Z_1 \\ \displaystyle\prod_{i=1}^{k-1} \left(\frac{n-1}{n-i+1}\right)^{\Lambda_i}, & Z_{k-1} \leq t < Z_k, \quad k = 2, \ldots, n, \\ 0, & t > Z_n. \end{cases}$$

Logo, o estimador de $F^o(t)$ é $\hat{F}_n = 1 - \hat{P}_n(t)$. Seja s_j a função salto de $\hat{P}_n$ em Z_j, definido por:

$$s_j = \begin{cases} 1 - \hat{P}_n(Z_2), & j = 1, \\ \hat{P}_n(Z_j) - \hat{P}_n(Z_{j+1}), & j = 2, \ldots, n-1 \\ \hat{P}_n(Z_n) & j = n. \end{cases} \tag{2.7}$$

Existem outros dois estimadores para a função de sobrevivência, o estimador da tabela de vida e o estimador de Nelson-Aalen, que não serão discutidos neste texto. Detatlhes sobre os mesmos poderão ser encontrados em Colosimo [2001].

2.2.3 Núcleo Estimador da Função de Densidade

Um estimador natural não-paramétrico de f^o, proposto por Marron e Padgett[1987], é definido por:

$$\begin{aligned} f_n(x) &= \frac{1}{h}\int_{-\infty}^{\infty} K\left(\frac{x-t}{h}\right) d\hat{F}_n(t) \\ &= \frac{1}{h}\sum_{j=1}^{n} s_j K\left(\frac{x-t}{h}\right), \end{aligned} \tag{2.8}$$

onde K é função núcleo.

Blum e Susarla [1980] propuseram um novo estimador não-paramétrico para a função de densidade através da função núcleo, a partir de resultados de Rosenblatt[1976] para dados censurados, motivados pelo fato que a estimativa de $f^o(x)H^*(x)$ é dada por

$$(f^o H^*)_n(x) = \frac{1}{nh}\sum_{j=1}^{n} K\left(\frac{x-Z_j}{h}\right) 1_{[\Lambda_j=1]}.$$

Para estimar $f^o(x)$, basta dividir $(f^o H^*)_n(x)$ por um estimador de $H^*(x)$. O estimador usado para $H^*(x)$, é o estimador produto-limite, dado por

$$\hat{H}_n(t) = \begin{cases} 1, & 0 \leq t \leq Z_1 \\ \displaystyle\prod_{i=1}^{k-1}\left(\frac{n-1}{n-i+1}\right)^{1-\Lambda_i}, & Z_{k-1} \leq t < Z_k, \quad k = 2,\ldots,n, \\ 0, & t > Z_n. \end{cases}$$

Consequentemente,

$$f_n^*(x) = \frac{\Lambda_j}{nh\hat{H}_n(Z_j)}\sum_{j=1}^{n} K\left(\frac{x-Z_j}{h}\right). \tag{2.9}$$

Observe, em (2.8) e (2.9), que

$$s_j = \frac{\Lambda_j}{n\hat{H}_n(Z_j)}.$$

Para medir a eficiência do estimador de $f^o(x)$, Marron e Padgett[1987] usaram o erro quadrático integrado $EQI(\hat{f}) = \int_0^\infty [\hat{f(x)} - f^o(x)]^2 dx$, devido às suas vantagens sobre o erro quadrático médio integrado citadas na Seção 1.2.

Para o teorema seguinte, assume-se que:

1. A função núcleo K é uma função de densidade de probabilidade com suporte compacto e Hölder contínua;
2. f^oH^* e f^o são Hölder contínuas de ordem $\alpha > 0$;
3. $h \to 0$ e $nh \to \infty$ quando $n \to \infty$;
4. F^o é absolutamente contínua;
5. H é contínua.

Teorema 2.2.1. *Sejam K, f^oH^*,F^o,f^o e H satisfazendo as condições $1-5$ e $h \in [n^{-1+\epsilon}, n^{-\epsilon}]$, onde ϵ é uma constante positiva tendendo a zero. Então,*

$$\sup_h \left| \frac{EQI(f_n) - EQMI(f_n)}{EQMI(f_n)} \right| \to 0$$

quase certamente.

A prova é encontrada em Padgett e Marron[1987].

Supondo que

$$\int x^j K(x)dx = \begin{cases} 1, & j = 0, \\ 0, & j = 1, \ldots, \mathsf{k}-1, \\ \kappa, & j = \mathsf{k} \end{cases}$$

e que f^o e f^oH^* têm k derivadas uniformementes contínuas, Marron e Padgett[1987] apresentam uma janela ótima h a partir do *Teorema 2.2.1.*

$$h_o = \left[\frac{(\int K^2) \int (f^o/H)}{2\mathsf{k}(\frac{\kappa}{\mathsf{k}!})^2 [\int ((f^o)^{(\mathsf{k})})^2]} \right]^{\frac{1}{2\mathsf{k}+1}} n^{-\frac{1}{2\mathsf{k}+1}} \tag{2.10}$$

com taxa de convergência do $EQI \approx n^{-\frac{2\mathsf{k}}{2\mathsf{k}+1}}$, onde $(f^o)^{(\mathsf{k})}$ denota a derivada de ordem k da função f^o.

2.2.4 Estimação da Janela Ótima

Desenvolvido para estimar a função densidade para amostras completas por Rudemo[1982] e Bowman[1984], o método da Validação Cruzada por Mínimos Quadrados foi adaptado para o caso onde a amostra é censurada à direita.

Desenvolvendo o $EQI(\hat{f})$, encontra-se

$$EQI(\hat{f}) = \int \hat{f}^2 - 2\int \hat{f}f^o + \int (f^o)^2.$$

O objetivo é encontrar a janela h que minimiza o $EQI(\hat{f})$. Observando a terceira parcela, vê-se que esta é independente de h. Então, procura-se o valor de h que minimiza a função escore

$$S(h) = \int \hat{f}^2 - 2\int \hat{f}f^o. \tag{2.11}$$

Como a primeira parcela é conhecida em (2.11), basta estimar a segunda parcela.

A integral da segunda parcela pode ser estimada por

$$\frac{1}{n}\sum_{i=1}^{n} \hat{f}_{n,i}(Z_i)\frac{1_{[\Lambda_i=1]}}{\hat{H}_n(Z_n)},$$

onde $\hat{f}_{n,i}$ é dada por

$$f_{n,i}(x) = \sum_{j\neq i} K\left(\frac{x - Z_j}{h}\right)\frac{1_{[\Lambda_j=1]}}{(n-1)\hat{H}_n(Z_j)h},$$

isto é, a função de densidade estimada em x utilizando-se todos os pontos amostrais exceto o ponto Z_i.

Portanto, para encontrar o valor da janela que minimiza o $EQI(f)$, basta encontrar o valor que minimiza a função a seguir:

$$CV(h) = \int [\hat{f}(x)]^2 dx - \frac{2}{n}\sum_{i=1}^{n} \hat{f}_i(Z_i)\frac{1_{[\Lambda_i=1]}}{\hat{H}(Z_i)}. \tag{2.12}$$

A equivalência na estimação de h entre (2.11) e (2.12) deve-se ao fato de que

$$E[CV(h)] = E[S(h)] \tag{2.13}$$

A igualdade em (2.13) pode ser verificada em Silverman[1986].

O resultado seguinte nos garantirá que o valor $\hat{h}_c$ que minimiza (2.12) é assintóticamente ótimo.

Teorema 2.2.2. *Sob as mesmas condições do Teorema 2.2.1, têm-se*

$$\frac{EQI(\hat{f}, \hat{h}_c)}{inf_h EQI(\hat{f}, \hat{h})} \to 1$$

quase certamente

A demonstração do *Teorema 2.2.2* é encontrada em Marron e Padgett[1987].

2.2.5 Núcleo Estimador da Função Quantílica

Para dados censurados à direita, um estimador natural para $Q^o(p)$ é o estimador baseado na função obtida usando o estimador **(PL)** da função distribuição F^o.

$$\tilde{Q}_n(p) = \inf\{t; \hat{F}_n(t) \geq p\}$$

Cheng[1984] obteve resultados de normalidade assintótica e Aly, *et al*[1985] apresentaram teoremas com grandes aproximações para o estimador da função quantílica através do produto limite $\tilde{Q}_n(p)$.

Observe que $\tilde{Q}_n(p)$ é uma função passo com saltos correspondentes às observações não-censuradas.

Com o objetivo de conseguir um estimador mais suave, Padgett[1986] e Lio, Padgett e Yu[1985], consideraram um estimador da função quantílica a partir da função núcleo, definindo-o por:

$$\begin{aligned} Q_n(p) &= \frac{1}{h_n}\int_0^1 \hat{Q}_n(p) K\left(\frac{t-p}{h_n}\right) dt \\ &= \frac{1}{h_n}\sum_{i=1}^{n} Z_i \int_{S_{i-1}}^{S_i} K\left(\frac{t-p}{h_n}\right) dt \end{aligned} \tag{2.14}$$

com $0 \leq p \leq 1$, sendo $S_i = \sum_{j=1}^{i} s_j$ com $i = 1, 2, \ldots, n$ e $S_0 = 0$, onde s_j é dado em (2.7).

Somente os $Z_i's$ correspondentes aos $X_i's$ não-censurados aparecem em (2.14) devido ao fato de que quando se tem censura, $S_i - S_{i-1} = 0$.

Resultados assintóticos para medir a eficiência de (2.14) são apresentados abaixo, fazendo antes algumas suposições.

1. Seja $\{h_n\}$ uma sequência de janelas positivas tal que $h_n \to 0$ quando $n \to \infty$;

2. Seja uma função real K definida em $(-\infty, \infty)$ tal que:

   (a) $K(x) \geq 0$ para todo número real x;

   (b) $\int_{-\infty}^{\infty} K(x)dx = 1$;

   (c) K tem suporte finito, isto é, $K(x) = 0$ para $|x| > c$ para alguma constante $c > 0$;

   (d) K é simétrica em torno de zero;

(e) K é uma função *Lipschitziana*.

3. F^o é contínua com densidade f^o;

4. f^o é contínua em $Q^o(p)$ e $f^o(Q^o(p)) > 0$, $0 < p < 1$;

5. $E(X^o) < \infty$,isto é, o valor esperado de F^o existe.

Definição 2.2.5. *Para uma função distribuição G, seja $T_G \equiv sup\{t : G(t) < 1\}$.*

Teorema 2.2.3. *Sob as condiçoes 1-5 e se $T_{F_o} < T_H \leq \infty$ e $h^{-1}(\log((\log n)/n))^{3/4} \to 0$ quando $n \to \infty$, então para cada $0 \leq p \leq 1$, $Q_n(p) \to Q^o(p)$ com probabilidade um, quando $n \to \infty$.*

Teorema 2.2.4. *Assuma as condições 1-5, que $T_{F_o} \leq T_H \leq \infty$ e que H é contínua em uma pequena vizinhança do intervalo aberto (T, T_o) de T_{F_o}. Seja ϕ uma função no intervalo fechado $[0, 1 - F(T)]$ tal que $\phi \geq x$, $\phi \to 0$ quando $x \to 0^+$ e $1 - F_o(t) \leq \phi(1 - F(t))$, para $t \in (T, T_{F_o})$. Se para qualquer $c > 1$, $\lim_n \sup \phi(c(\log\log n/(2n))^{1/2})h^{-1} < \infty$, então para cada $0 \leq p \leq 1$, $Q_n(p) \to Q^o(p)$ com probabilidade um, quando $n \to \infty$.*

Como na prática as amostras são pequenas, criou-se um enorme problema. Nenhuma propriedade para amostra pequena foi deduzida. Na realidade, devido às complicações matemáticas introduzidas pela censura, uma expressão exata para o erro quadrado medio de $Q_n(p)$ para n pequeno não está disponível. Assim, um valor da janela que minimiza o erro quadrado médio exato de $Q_n(p)$ não pode ser obtido. Um bom método prático para selecionar a janela h é o método do bootstrap, descrito na próxima seção.

2.2.6 Seleção da Janela Usando o Método do Bootstrap

Na prática em geral, não é conhecido a forma da distribuição dos dados e um estimador não paramétrico é proposto. Consequentemente, o estimador $Q_n(p)$ é um estimador não paramétrico e deseja-se um método para selecionar a janela ótima que não dependa do conhecimento da forma da distribuição, isto é, dada uma amostra aleatória censurada de tamanho n, qual é a janela h ótima para o cálculo de $Q_n(p)$? O

método do Bootstrap para dados censurados provê uma solução a este problema para o critério do mínimo do erro quadrático médio. Este método estima o erro quadrático médio de $Q_n(p)$, o $EQM(Q_n(p))$, como uma função de h e escolhe o valor de janela h que minimiza este erro.

Da amostra censurada inicial, denotada por (Z_i, Λ_i), $i = 1, \ldots, n$, o método do Bootstrap consiste em retiradas aleatórias com reposição e com probabilidade $\frac{1}{n}$ de novas amostras (Z_i^*, Λ_i^*), $i = 1, ..., n$, a partir das observações bivariadas (Z_i, Λ_i), $i = 1, ...n$, isto é, retira-se com probabilidade $\frac{1}{n}$ uma observação Z_j acompanhado do seu respectivo Λ_j, denotando-os agora por (Z_i^*, Λ_i^*), $i = 1, ..., n$.

Denote o estimador produto limite de $Q^o(p)$ baseado nas amostras bootstrap por $\tilde{Q}_n^*(p)$ e seja $Q_n^*(p)$ o estimador através do núcleo baseado também nas amostras bootstrap, isto é,

$$Q_n^*(p) = \frac{1}{h} \int_0^1 \hat{Q}_n^*(t) K\left(\frac{t-p}{h}\right) dt.$$

O erro quadrático médio EQM pode ser decomposto na soma da variância e vício ao quadrado:

$$EQM() = Var() + [Vício()]^2.$$

O estimador da variância de $Q_n(p)$ é dado por:

$$\hat{Var}(h) = \frac{1}{(B-1)} \left\{ \sum_{i=1}^{B} |Q_{ni}^*(p)|^2 - \frac{|\sum_{i=1}^{B} Q_{ni}(p)|^2}{B} \right\}$$

e o estimador do vício é

$$\hat{Vício}_1(h) = \frac{1}{B} \sum_{i=1}^{B} Q_{ni}^*(p) - Q_n(p).$$

Seja $EQM_{Q_n(p)}^*$ o erro quadrático médio do estimador bootstrap de $Q_n(p)$.

Padgett e Thombs[1986] observaram em várias simulações que, fixando o percentil p e variando o valor de h de forma crescente, o quadrado do vício aumentava enquanto a variância diminuia. Desta forma, $EQM_{Q_n(p)}(h)$, como função de h, deveria decrescer e depois crescer, e a estimativa $EQM_{Q_n(p)}^*(h)$ daria o valor de h que minimizaria o $EQM_{Q_n(p)}(h)$.

Entretanto, em muitas situações encontradas nas simulações em Pagett[1986], $EQM_{Q_n(p)}^*(h)$ era estritamente decrescente em h. O motivo seria devido a ambos

$Q_n(p)$ e o estimador da amostra bootstrap $Q_n^*(p)$ serem super suavizados, resultando numa estimativa pobre do vício.

Para contornar este problema, trocou-se o estimador quantílico do núcleo sem bootstrap $Q_n(p)$ pelo estimador quantílico produto limite $\tilde{Q}_n(p)$ que não depende da janela h, na estimativa do vício, isto é,

$$\hat{V}ício_2(h) = \frac{1}{B}\sum_{i=1}^{B} Q_{ni}^*(p) - \tilde{Q}_n(p),$$

e consequentemente,

$$EQM_{Q_n(p)}^*(h) = \hat{V}ar(h) + \hat{V}ício_2^2(h).$$

Para selecionar a melhor janela e comparar qual é o melhor estimador para a função quantílica através do método do Bootstrap, seguiu-se a recomendação feita por Padgett[1986], onde é calculada a razão entre o erro quadrático médio do estimador quantílico produto-limite e o erro quadrático médio do estimador quantílico através da função núcleo, para uma sequência de janelas variando de $0,01$ a $0,60$ acrescidos de $0,01$. Escolhe-se a janela que apresentou maior razão. Se este valor for superior a um tem-se que o erro quadrático médio do estimador quantílico através da função núcleo é menor que o erro quadrático médio do estimador quantílico produto-limite. Em seguida, calculou-se a função quantílica para um percentil fixo.

2.2.7 Estimação Intervalar Usando Bootstrap

Todos os estimadores para o intervalo de confiança assumem que a janela h é conhecida e o percentil fixo. O primeiro estimador intervalar proposto por Thombs e Padgett[1986], chamado nesta dissertação de *Intervalo* 1, é definido por:

$$\left[Q_n(p) - Z_{(1-\alpha/2)}\sqrt{\hat{V}ar(h)}, Q_n(p) + Z_{(1-\alpha/2)}\sqrt{\hat{V}ar(h)}\right]$$

onde $Z_{1-\alpha/2}$ é o percentil $(1-\alpha/2)$ da distribuição normal padrão e $\hat{V}ar(h)$ é a variância estimada usando bootstrap calculado da seguinte forma:

$$\hat{V}ar(h) = \frac{1}{(B-1)}\left\{\sum_{i=1}^{B}|Q_{ni}^*(p)|^2 - \frac{|\sum_{i=1}^{B}Q_{ni}(p)|^2}{B}\right\}.$$

O segundo intervalo, *Intervalo 2*, também proposto por Padgett e Thombs[1986] para estimar intervalos de confiança para a função quantílica é obtido usando-se as estimativas bootstrap de $Q_n(p)$ e construindo a sua distribuição empírica

$$G(t) = P^*(Q_n(p) \leq t) = \frac{\#\{Q^*_{ni}(p) \leq t\}}{B},$$

onde $Q^*_{ni}(p)$ é o quantil de ordem p da i-ésima amostra bootstrap, B é o número de amostras bootstrap obtidas, onde o valor sugerido por Padgett e Thombs[1986] é de 1000 reamostragens. Nesta dissertação foram realizadas 300 reamonstragens para efeito de teste.

Então, a aproximação para o intervalo de confiança para $Q^0(p)$ é dada por:

$$\left[G^{-1}\left(\frac{\alpha}{2}\right), G^{-1}\left(1-\frac{\alpha}{2}\right)\right]$$

O último intervalo, denominado nessa dissertação por *Intervalo 3*, para estimar o intervalo de confiança a ser avaliado é o método apresentado no artigo do Cheng[2002].

$$\left[Q_n(p) - Z_{(1-\alpha/2)}\sqrt{\frac{q_n^2(p)p(1-p)}{n}}, Q_n(p) + Z_{(1-\alpha/2)}\sqrt{\frac{q_n^2(p)p(1-p)}{n}}\right]$$

onde $q_n(p)$ é a derivada do estimador quantílico $Q_n(p)$, e é dado por:

$$q_n(p) = \frac{1}{h}\sum Z_i \int_{S_{i-1}}^{S_i} K'(\frac{t-p}{h})dt.$$

O grande mérito aqui é estimar a quantidade $q_n(p)$. Primeiramente, utiliza-se a janela selecionada na estimação da função quantílica e calcula-se a derivada quantílica. Os resultados apresentados foram bem distantes dos valores teóricos realizados nas simulações. Um segundo processo é usar a janela h obtida para $Q_n(p)$ e calcular as estimativas para a derivada $q_n(p)$ através do limite

$$q_n(p) = \lim_{\delta_j \to 0} \frac{Q_n(p+\delta_j) - Q_n(p)}{\delta_j}.$$

Este é o processo numérico para diferenciação apresentado em Mathews[1987], tomando uma sequência δ_j de tal modo que $\delta_j \to 0$:

$$q_n(p) = \frac{Q_n(p+\delta_j) - Q(p)}{\delta_j}.$$

No trabalho $\delta_1 = 10^{-2}$, $\delta_2 = 10^{-3}$, $\delta_3 = 10^{-4}$, ..., $\delta_j = 10^{-(j+1)}$,...

Trabalhou-se com as seguintes aproximações para $q_n(p)$:

1. Derivada a direita:

$$q_{n1}(p) = \frac{Q_n(p+\delta_j) - Q_n(p)}{\delta_j};$$

2. Derivada a esquerda:

$$q_{n2}(p) = \frac{Q_n(p) - Q_n(p-\delta_j)}{\delta_j};$$

3. Derivada central:

$$q_{n3}(p) = \frac{Q_n(p+\delta_j) - Q_n(p-\delta_j)}{2\delta_j};$$

4. Derivada com quatro pontos:

$$q_{n4}(p) = \frac{-Q_n(p+2\delta_j) + 8Q_n(p+\delta_j) - 8Q_n(p-\delta_j) + Q_n(p-2\delta_j)}{12\delta_j},$$

Capítulo 3

Simulações e Resultados

3.1 Introdução

As simulações têm como objetivo testar o desempenho dos métodos para a estimação em estudo comparando-os com seus valores teóricos. Através de simulações, buscou-se avaliar o desempenho do método do núcleo na seleção da janela h e assim na estimativas pontuais e intervalares.

Para efeito de simulação, usou-se as distribuições Weibull e Lognormal devido a estes modelos serem os que mais bem se ajustam em aplicações reais em análise de sobrevivência. Considere as seguintes situações:

1. Tempo de vida Weibull com parâmetros $\lambda = 6$(escala) e $\beta = 6$(forma) e para censura $\lambda = 6$ e $\beta = 7$, resultando em 28% de censura.

2. Tempo de vida Logmormal com parâmetros $\mu = 1,5$ e $\sigma = 0,14$, e para censura uma Lognormal com parâmetros $\mu = 1,6$ e $\sigma = 0,14$, resultando em 30% de censura.

Primeiramente, usou-se as distribuições apresentadas em 1 gerando amostras de tamanho n=50 para estimar a janela h para a função densidade usando o método da Validação Cruzada. Já na segunda parte, trabalhou-se com as distribuições dos dois itens citados com amostras de tamanho 30, 50, 100 e 300. A cada amostra gerada, realizou-se 300 reamostragens para determinar a janela h para estimar a função quantílica.

3.2 Implementação

Para a implementação computacional, foi utilizado o *software dev C++* devido a sua liberdade de programação e velocidade na execução das simulações. Foram geradas amostras selecionadas aleatoriamente das populações pré-especificadas.

Para gerar números aleatórios (pseudo-aleatório), foi usado o próprio gerador do dev C++, onde a função *rand*() fornece números inteiros no intervalo de 0 a RAND_MAX, onde RAND_MAX é uma constante definida na biblioteca <*cstdlib*> do C++. Para gerar amostras uniformes $U(0,1)$, simplesmente dividiu-se o número aleatório gerado pelo valor RAND_MAX. Em seguida, utilizou-se o método da função inversa para gerar amostras da distribuição Weibull, e para a Lognormal, usou-se o método Box-Muller para geração de normais(X), seguido da transformação $Y = exp(X)$.

O tempo de execução das 100 simulações realizadas ficou em média em seis minutos, cálculos realizados em um computador PENTIUN III de 1 Gb com mémoria de 256 Kb. A parte gráfica foi realizada no ambiente de programação R na versão 2.0.1.

O algoritmo para os cálculos das estimativas encontra-se no Apêndice C.

Medidas de desempenho, como erro quadrático médio (2.1) e erro quadrático médio integrado (2.2), foram calculadas para viabilizar as comparações das estimativas da função densidade e da função quantil.

3.3 Resultados

As simulações, tanto na estimação pontual quanto na intervalar, foram realizadas com o objetivo de comparar e conhecer vantangens e desvantagens do método de estimação não paramétrico.

3.3.1 Função de Densidade

Das cem simulações realizadas, para amostras de tamanho n=50, resultaram cem janelas selecionadas através do método da Validação Cruzada. Tomou-se a média destas janelas obtendo o valor $h = 0,52$ e calculou-se a janela ótima apresentada em

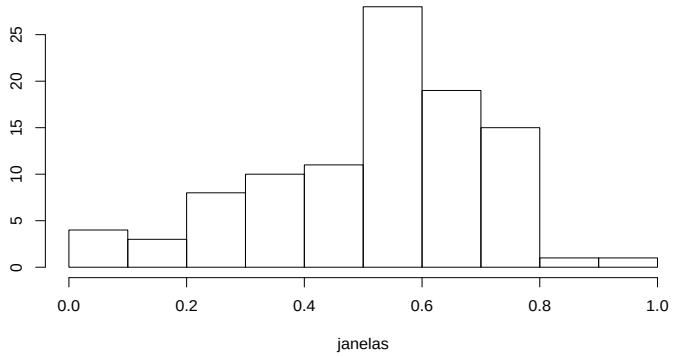

Figura 3.1: Histograma para as janelas selecionadas das 100 simulações

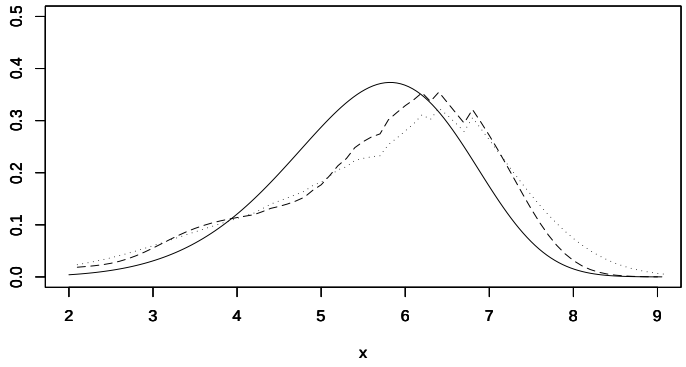

Figura 3.2: função de Densidade

Padgett e Marron[1986], encontrando o valor $h_{opt} = 0,82$.

A Figura 3.1 é o histograma das janelas selecionadas pelo método da Validação Cruzada para as 100 simulações realizadas. O histograma é assimétrico e existe grande variabilidade nas janelas selecionadas.

A Figura 3.2 mostra a forma da função de densidade usando $h = 0,52$(*linha tracejada*), $h_{opt} = 0,82$(*linha pontilhada*) e a curva teórica da Weibull(6,6)(*linha contínua*).

A média da janela estimada subestima a janela teórica (2.10) apresentada por Padgett[1987]. Foi usado apenas uma única amostra.

3.3.2 Função Quantil

Estimador Pontual

Para selecionar a janela para a função quantílica, foi usado o método computacional Bootstrap com 300 reamostragens (sugerido por Padgett e Thombs[1986]). Para cada uma das 100 amostras geradas (100 simulações), selecionou-se a janela h resultante do menor erro quadrático médio para o bootstrap. Lembrando que as janelas eram escolhidas no intervalo $[0,01; 0,60]$ com acréscimo de $0,01$.

Utilizou-se a média aritmética das janelas($\overline{h}_{Boot}$) selecionadas em cada simulação e calculou-se os quantis $0,1$; $0,5$; $0,75$ e $0,95$ para o estimador quantílico para as distribuições já descritas comparando-os com seus respectivos valores teóricos. Os resultados podem ser vistos nas Tabelas 3.1 e 3.3, onde $dp_{\overline{h}_{Boot}}$ é o desvio padrão de $\overline{h}_{Boot}$. Construiu-se alguns histogramas para as janelas selecionadas.

Tabela 3.1: Estimativas pontuais dos quantis para Distribuição Weibull

Quantil	Amostra	$\overline{h}_{Boot}$	$dp_{\overline{h}_{Boot}}$	$Q_{\overline{h}_{Boot}}(p)$	$Q^o(p)$
0,1	30	0,1436	0,0717	4,2411	4,1234
	50	0,1035	0,0407	3,9711	
	100	0,0832	0,0411	4,2207	
	300	0,0540	0,0334	4,1443	
0,5	30	0,3614	0,1653	5,9497	5,6445
	50	0,2805	0,1920	5,9962	
	100	0,3104	0,1880	5,5844	
	300	0,2668	0,1850	5,6888	
0,75	30	0,2660	0,0638	6,8018	6,3357
	50	0,2739	0,0564	6,6001	
	100	0,2230	0,1021	6,3784	
	300	0,2328	0,0993	6,3097	
0,95	30	0,0242	0,0161	7,5702	7,2039
	50	0,0403	0,0143	7,2384	
	100	0,0501	0,0010	6,9950	
	300	0,0492	0,0120	7,2089	

Na Figura 3.3, observa-se algumas janelas próximas de $0,01$ ou próximas de $0,5$, um pouco distantes da média $0,1436$. Para algumas destas janelas, as estimativas

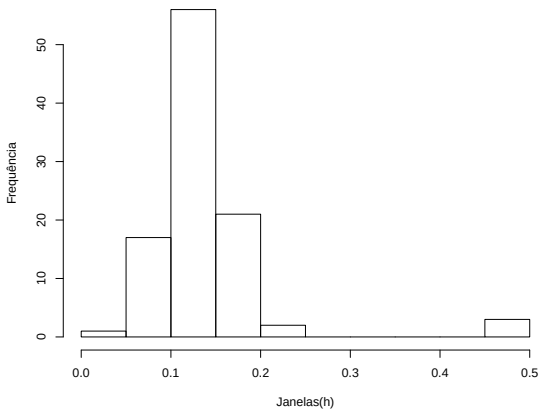

Figura 3.3: Hist; p=0,1; n=30; WB

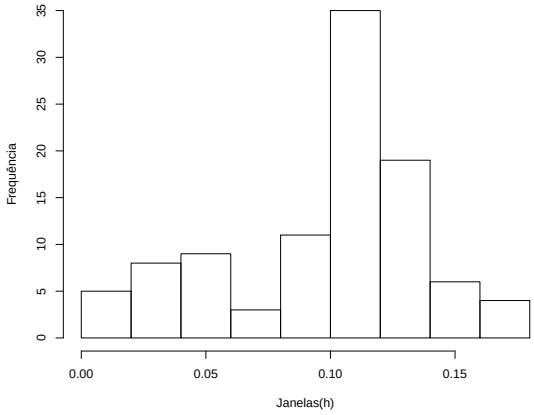

Figura 3.4: Hist; p=0,1; n=50; WB

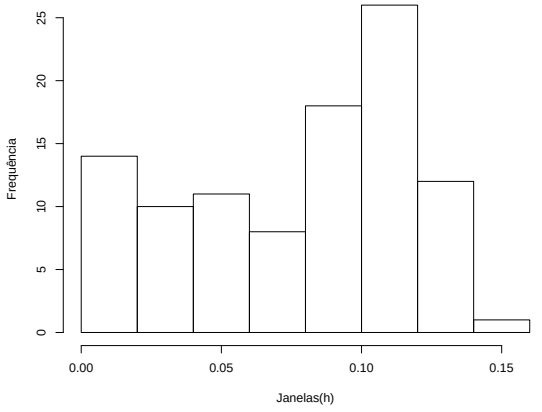

Figura 3.5: Hist; p=0,1; n=100; WB

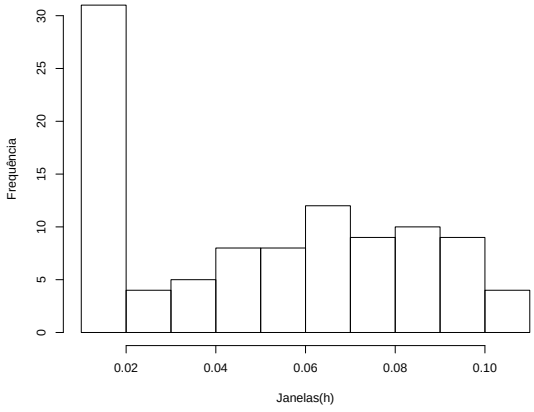

Figura 3.6: Hist; p=0,1; n=300; WB

Tabela 3.2: Estimativas pontuais dos quantis para Distribuição Weibull

Janela	0,01	0,01	0,06	0,06	0,06	0,07	0,5	0,5	0,5
$Q_n(0,1)$	4,3136	4,3136	4,4037	4,5061	4,3913	4,0481	2,9922	2,9348	3,0875

Tabela 3.3: Estimativas pontuais dos quantis para Distribuição Lognormal

Quantil	Amostra	$\overline{h}_{Boot}$	$dp_{\overline{h}_{Boot}}$	$Q_{\overline{h}_{Boot}}(p)$	$Q^o(p)$
0,1	30	0,1174	0,0263	3,5456	3,7464
	50	0,0965	0,0344	3,6499	
	100	0,0891	0,0332	3,8957	
	300	0,0606	0,0378	3,6581	
0,5	30	0,4796	0,1582	4,4321	4,4817
	50	0,4378	0,1954	4,4161	
	100	0,3911	0,2102	4,5604	
	300	0,3960	0,2150	4,4580	
0,75	30	0,2628	0,0678	4,7536	4,9224
	50	0,2551	0,0802	4,7853	
	100	0,2342	0,1047	5,0508	
	300	0,2153	0,1117	4,9427	
0,95	30	0,0233	0,0154	5,2489	5,6384
	50	0,0340	0,0143	5,2511	
	100	0,0452	0,0131	5,8243	
	300	0,0489	0,0126	5,9672	

para $Q(0,1)$ foram boas, outras apresentaram estimativas bem ruins, como pode ser observado na Tabela 3.2. Lembrando que o valor teórico para o percentil $0,1$ é de $4,1234$.

Quando passou-se para amostras de tamanho $n = 300$, a maioria das janelas estimadas ficaram entre 0 e $0,02$, como pode ser observado na Figura 3.6. Mas, mesmo com estas janelas, as estimativas foram boas, como pode ser visto na Tabela 3.1. Para os demais percentis o comportamento foi parecido.

Também foi feito o mesmo estudo para a distribuiçao Lognormal (Tabela 3.3 e Figuras 3.7 a 3.10)

Os resultados das Tabelas 3.1 e 3.3 mostram que, usando a média aritmética das janelas das cem simulações realizadas, as estimativas pontuais estão próximas dos

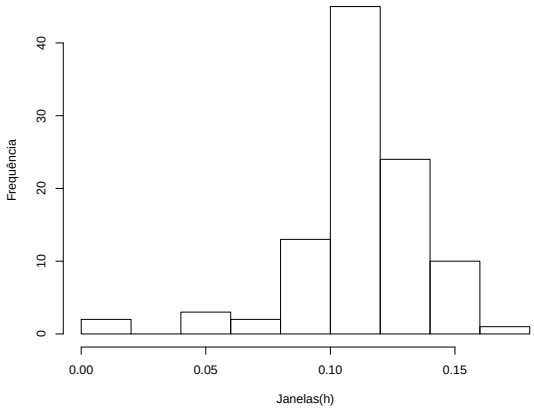

Figura 3.7: Hist; p=0,1; n=30; LN

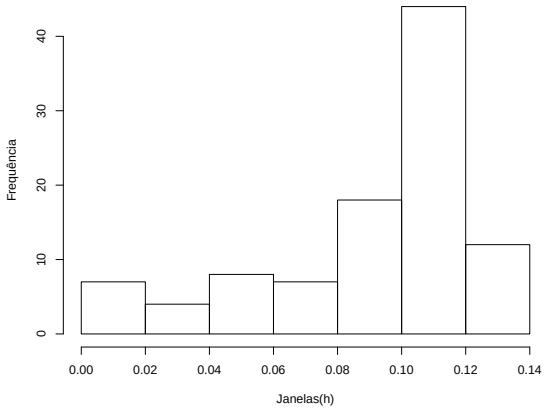

Figura 3.8: Hist; p=0,1; n=50; LN

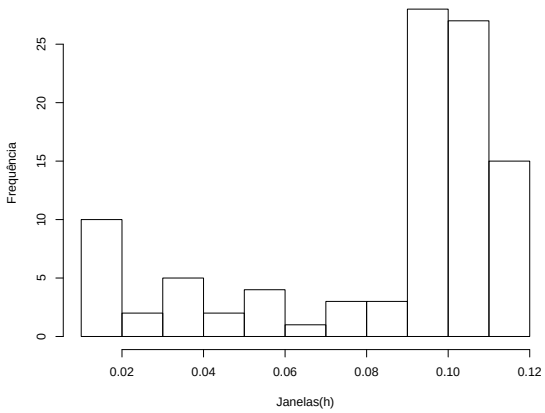

Figura 3.9: Hist; p=0,1; n=100; LN

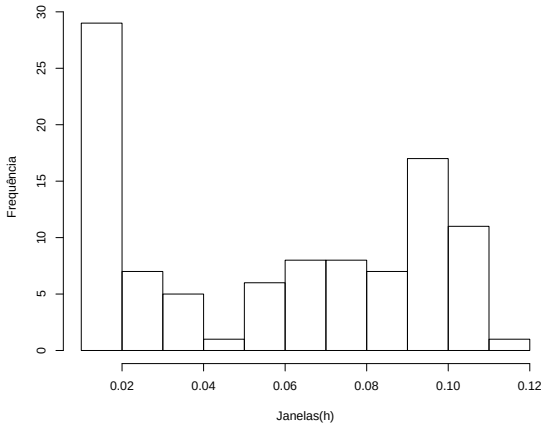

Figura 3.10: Hist; p=0,1; n=300; LN

valores teóricos. Também nota-se que, com o aumento no tamanho das amostras, há na melhoria das estimativas.

Estimador Intervalar

Ao contrário do que foi feito com as estimativas pontuais, onde usou-se a média aritmética das cem janelas selecionadas, para calcular os cem intervalos, usou-se o estimador pontual com sua respectiva janela selecionada, isto é, para cada intervalo calculado foi usada a janela selecionada através do bootstrap. Lembrando que foram três tipos de intervalos descritos na seção 2, sendo que para o intervalo denominado *intervalo 3*, há necessidade de estimar a derivada da função quantílica.

As Tabelas 3.4 a 3.6 apresentam as aproximações para a derivada $q_n(p)$, lembrando que as estimativas foram feitas com o valor da janela h selecionada através da técnica do bootstrap para a estimativa do quantil $0,1$. Como o objetivo é estimar a derivada para a função quantil $q_n(p)$, para efeito de teste, estimou-se $q^o(p)$ usando a distribuição Weibull(6;6) cujo valor teórico da derivada no ponto $0,1$ é $7,2475$. As estimativas foram feitas com amostras de tamanhos 30, 50 e 100.

Para cada uma das cinco janelas selecionadas, observa-se que a melhor aproximação foi para derivada à direita com $\delta = 0,01$. O mesmo ocorreu para demais quantis e tamanhos de amostra. Portanto, adotou-se esta aproximação para o cálculo do *Intervalo 3*. Um fato interessante é que as janelas $0,1$ e $0,11$ apresentaram melhores resultados para o estimador quantílico e para a aproximação para a derivada, janelas um pouco distantes da média usada na estimação pontual que foi de $0,1436$.

Um comparativo entre os três métodos de estimação intervalar foi realizado com cem intervalos gerados com 90% de confiança. As Tabelas 3.7 a 3.10 apresentam o número de intervalos que contiveram o verdadeiro valor quantil (valor teórico) para a distribuição Weibull e Lognormal para amostras de tamanhos 30, 50, 100 e 300, e quantis de $0,1$, $0,5$, $0,75$ e $0,95$.

Nas Tabelas 3.7 e 3.9, vê-se que o *Intervalo 3* para o quantil $0,1$ apresenta melhores resultados do que os *Intervalos 1 e 2*, isto é, existe um número maior de intervalos que contiveram o verdadeiro parâmetro. Para a mediana $(0,5)$, nenhum dos três intervalos ganhou destaque.

Tabela 3.4: Estimativas para $q(0,1)$; n=30

Janela(h)	Q_n	δ_j	$q_{n1}(p)$	$q_{n2}(p)$	$q_{n3}(p)$	$q_{n4}(p)$
0,13	3,9681	0,01	9,84209	12,0295	10,9358	10,9527
		0,001	10,8598	11,0725	10,9662	10,9662
		0,0001	10,9555	10,9768	10,9662	10,9662
		1e-05	10,9651	10,9672	10,9662	10,9662
0,11	4,12293	0,01	7,42305	10,5087	8,96588	8,86883
		0,001	8,81363	9,12182	8,96772	8,96772
		0,0001	8,95231	8,98313	8,96772	8,96772
		1e-05	8,96618	8,96926	8,96772	8,96772
0,1	4,23995	0,01	7,92792	10,4967	9,21230	8,97262
		0,001	8,44081	8,69831	8,56956	8,54587
		0,0001	8,49272	8,51847	8,50559	8,50323
		1e-05	8,49791	8,50049	8,49920	8,49896
0,18	3,65764	0,01	13,2513	14,1204	13,6859	13,6972
		0,001	13,6855	13,7721	13,7288	13,7305
		0,0001	13,7289	13,7376	13,7333	13,7334
		1e-05	13,7333	13,7342	13,7337	13,7337
0,15	4,0847	0,01	13,7566	15,4475	14,6021	14,6376
		0,001	14,6144	14,7835	14,6990	14,7026
		0,0001	14,7002	14,7171	14,7087	14,7090
		1e-05	14,7088	14,7105	14,7096	14,7097

Tabela 3.5: Estimativas para $q(0,1)$; n=50

Janela(h)	Q_n	δ_j	$q_{n1}(p)$	$q_{n2}(p)$	$q_{n3}(p)$	$q_{n4}(p)$
0.12	3.8290	0.01	12.5509	14.6887	13.6198	13.6337
		0.001	13.5415	13.7498	13.6456	13.6456
		0.0001	13.6352	13.656	13.6456	13.6456
		1e-005	13.6446	13.6467	13.6456	13.6456
0.09	3.87747	0.01	14.5702	16.0509	15.3106	15.2892
		0.001	15.6468	15.7935	15.7202	15.7356
		0.0001	15.7545	15.7692	15.7618	15.7634
		1e-005	15.7653	15.7667	15.766	15.7661
0.13	3.78502	0.01	14.0748	15.8346	14.9547	14.9652
		0.001	14.8996	15.0696	14.9846	14.9846
		0.0001	14.9761	14.9931	14.9846	14.9846
		1e-005	14.9837	14.9854	14.9846	14.9846
0.13	3.96042	0.01	16.6135	18.4479	17.5307	17.5467
		0.001	17.4956	17.6687	17.5822	17.5827
		0.0001	17.5735	17.5908	17.5822	17.5822
		1e-005	17.5813	17.583	17.5822	17.5822
0.04	4.81367	0.01	3.07893	2.77529	2.92711	2.93772
		0.001	2.94966	2.91796	2.93381	2.93378
		0.0001	2.93539	2.93222	2.93381	2.93381
		1e-005	2.93397	2.93365	2.93381	2.93381

Tabela 3.6: Estimativas para $q(0,1)$; n=100

Janela(h)	Q_n	δ_j	$q_{n1}(p)$	$q_{n2}(p)$	$q_{n3}(p)$	$q_{n4}(p)$
0,12	3,9425	0,01	11,7995	14,0482	12,9238	12,951
		0,001	12,8611	13,0757	12,9684	12,9686
		0,0001	12,9576	12,9791	12,9684	12,9684
		1e-005	12,9673	12,9694	12,9684	12,9684
0,09	3,7844	0,01	9,43993	10,4889	9,96443	9,79656
		0,001	9,61936	9,73174	9,67555	9,66469
		0,0001	9,64084	9,65736	9,6491	9,64856
		1e-005	9,64827	9,64993	9,6491	9,6491
0,01	4,0537	0,01	3,09002	2,7644	2,92721	2,48412
		0,001	1,18828	1,47358	1,33093	1,30929
		0,0001	1,31425	1,34322	1,32873	1,32873
		1e-005	1,32728	1,33018	1,32873	1,32873
0,06	4,0017	0,01	7,45174	7,931	7,69137	7,65355
		0,001	7,62456	7,67209	7,64832	7,64709
		0,0001	7,64392	7,64904	7,64648	7,64648
		1e-005	7,64623	7,64674	7,64648	7,64648
0,09	4,0471	0,01	7,19484	8,74264	7,96874	7,81528
		0,001	7,76751	7,90409	7,8358	7,82746
		0,0001	7,80632	7,81998	7,81315	7,81232
		1e-005	7,81021	7,81157	7,81089	7,81081

Tabela 3.7: Número de intervalos dist. Weibull

	Quantis							
	0.1				0.5			
n	30	50	100	300	30	50	100	300
Intervalo 1	73	78	76	79	88	80	89	88
Intervalo 2	78	79	82	84	93	81	88	87
Intervalo 3	93	85	83	84	92	82	84	88

Tabela 3.8: Número de intervalos dist. Weibull

	Quantis							
	0.75				0.95			
n	30	50	100	300	30	50	100	300
Intervalo 1	89	77	82	86	57	71	71	74
Intervalo 2	88	80	84	86	36	48	66	79

Tabela 3.9: Número de intervalos dist. Lognormal

	Quantis							
	0.1				0.5			
n	30	50	100	300	30	50	100	300
Intervalo 1	81	86	91	89	86	82	86	82
Intervalo 2	87	88	91	90	91	80	89	84
Intervalo 3	99	94	94	87	86	83	83	87

Tabela 3.10: Número de intervalos dist. Lognormal

	Quantis							
	0.75				0.95			
n	30	50	100	300	30	50	100	300
Intervalo 1	87	78	85	83	54	61	69	66
Intervalo 2	88	81	88	82	27	37	58	79

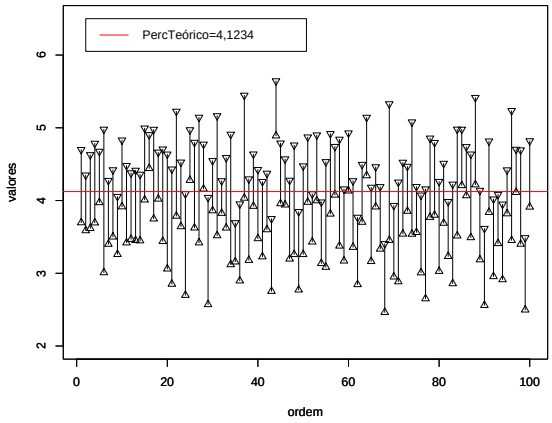

Figura 3.11: Int1; p=0,1; n=30; WB

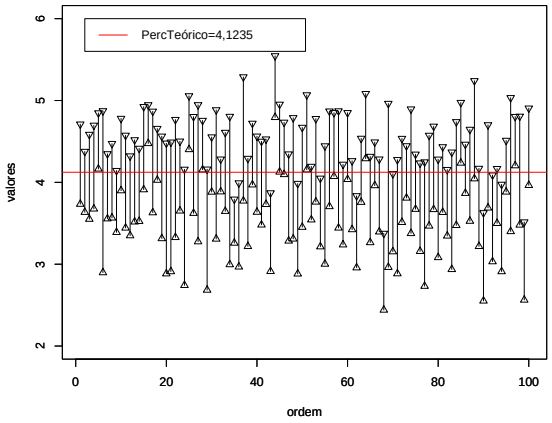

Figura 3.12: Int2; p=0,1; n=30; WB

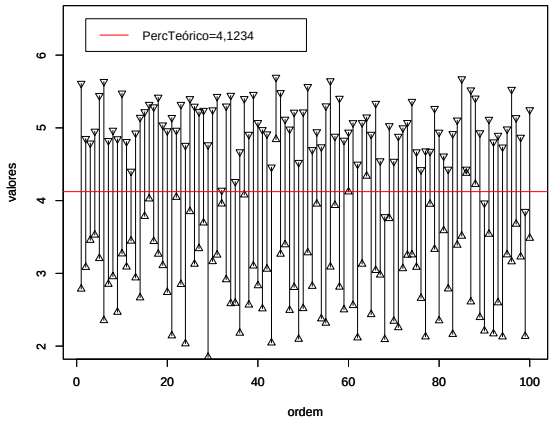

Figura 3.13: Int3; p=0,1; n=30; WB

Um fato interessante é que, para quantis altos, Tabelas 3.8 e 3.10, o *Intervalo 3* apresentou-se de forma irregular tornando-se inviável. Mesmo para os *Intervalos 1 e 2*, a quantidade de intervalos que contiveram o verdadeiro parâmetro foi muito baixa.

Através destes resultados, pode-se concluir que, para quantis superiores à mediana, estes três métodos de estimação intervalar não têm bom desempenho.

As Figuras 3.11 a retratam os três intervalos para o quantil $0,1$ e $0,5$ com o tamanho da amostra $n = 30$ para as distribuições Weibull $WB(6,6)$ e Lognormal $LN(1,5;0,0196)$. A notação Int1 significa *Intervalo 1*, de maneira análoga para Int2 e Int3. Os demais casos poderão ser vistos no *Apêndice A*.

Nas Figuras 3.11 a 3.13, percebeu-se que o *Intervalo 3* apresenta amplitude maior

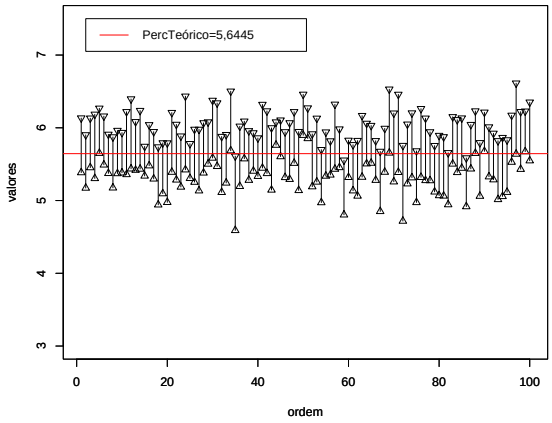

Figura 3.14: Int1; p=0,5; n=30; WB

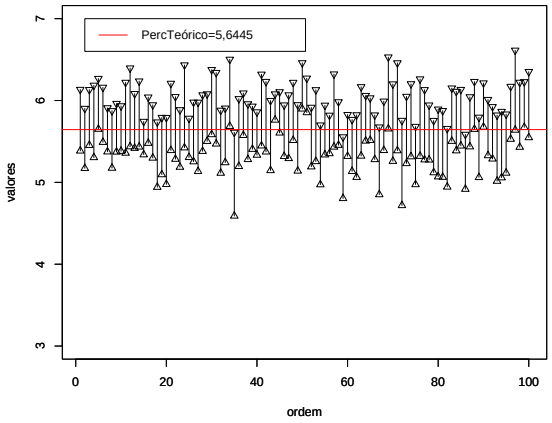

Figura 3.15: Int2; p=0,5; n=30; WB

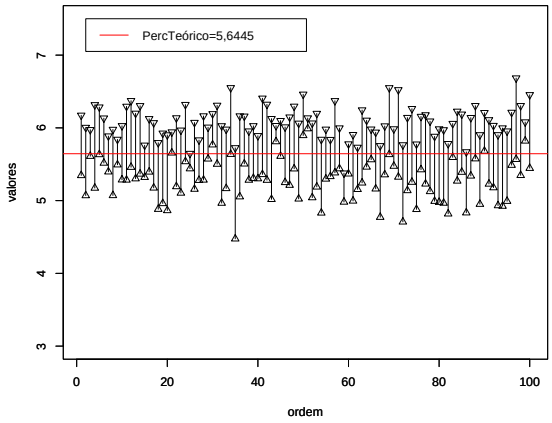

Figura 3.16: Int3; p=0,5; n=30; WB

do que os demais intervalos, fato este que contribuiu para um maior número de vezes contendo o verdadeiro parâmetro.

Nas Figuras 3.14 a 3.16, os três intervalos se comportaram de maneira similar, sem destaque para nenhum intervalo.

As Figuras 3.17 a 3.22, onde a distribuição usada foi a Lognormal, apresentaram o mesmo comportamento das Figuras 3.11 a 3.16.

Continuando a avaliação dos estimadores através do método núcleo, em Padgett e Thombs[1986] encontra-se um conjunto de dados censurados gerados de duas distribuições exponenciais, a primeira com parâmetro $\lambda = 1$ para tempo de vida e a segunda com parâmetro $\lambda = 3/7$ para censura, resultando em 30% de censura.

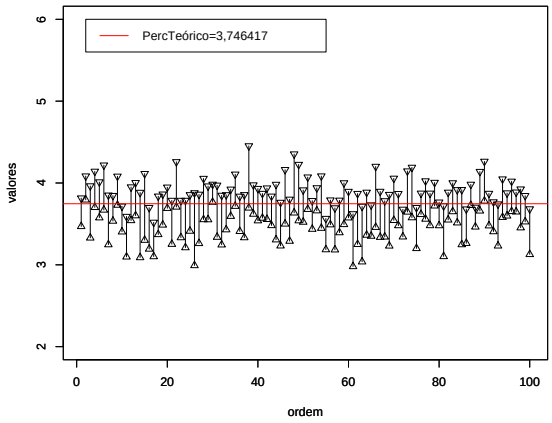

Figura 3.17: Int1; p=0,1; n=30; LN

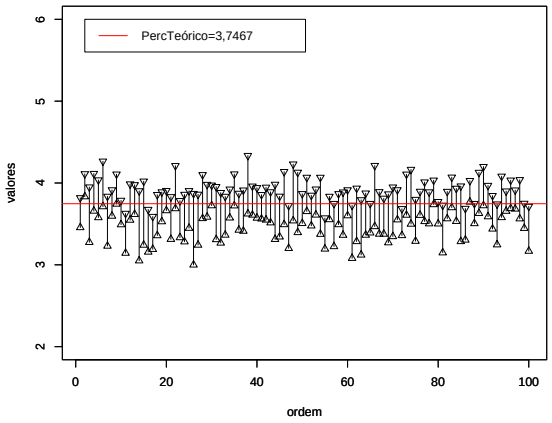

Figura 3.18: Int2; p=0,1; n=30; LN

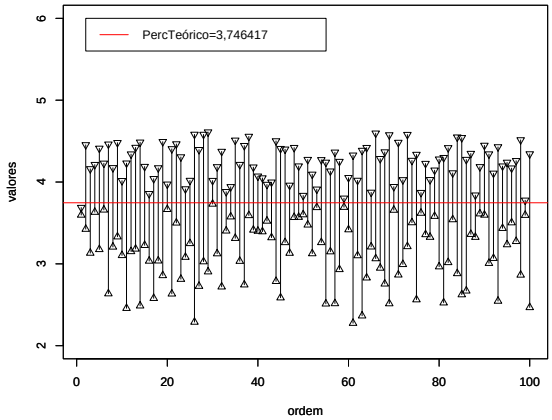

Figura 3.19: Int3; p=0,1; n=30; LN

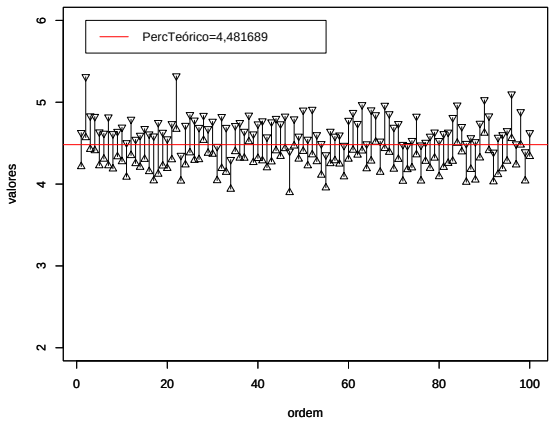

Figura 3.20: Int1; p=0,5; n=30; LN

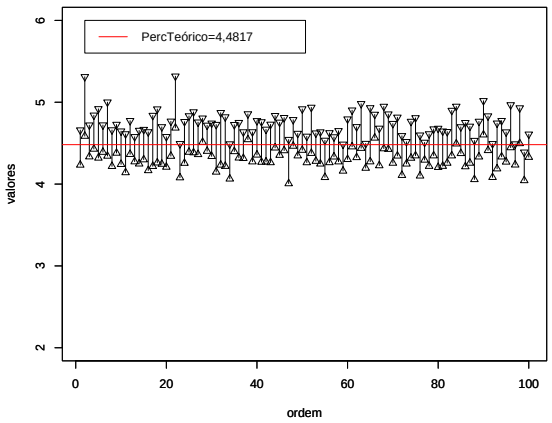

Figura 3.21: Int2; p=0,5; n=30; LN

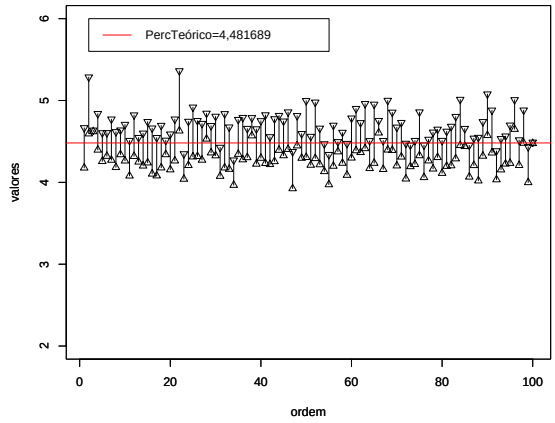

Figura 3.22: Int3; p=0,5; n=30; LN

Tabela 3.11: Estimativas pontuais do artigo Padgett e Thombs[1986]

Quantil	$Q^0(p)$	h	$Q_n(p)$	*Intervalo 1*	*Intervalo 2*
0,10	0,1054	0,17	0,1462	(0,0849; 0,2076)	(0,0953; 0,2194)
0,50	0,6931	0,23	0,7110	(0,5781; 0,9099)	(0,5903; 0,9118)
0,75	1,3863	0,49	1,2605	(0,9466; 1,5744)	(0,9735; 1,5862)

A Tabela 3.11 refere-se às estimativas apresentadas no artigo destes autores.

Usando o mesmo conjunto de dados apresentado em Padgett e Thombs[1986], conseguiu-se os resultados da Tabela 3.12.

Num comparativo entre as Tabelas 3.11 e 3.12 para as quantidades comuns, percebe-se que, para as estimativas pontuais dos quantis, os valores da Tabela 3.11 estão melhores do que os da Tabela 3.12, isto é, os valores da Tabela 3.11 estão mais

Tabela 3.12: Estimativas dos quantis para dist Exp(1)

Quantil	$Q^0(p)$	h	$Q_n(p)$	*Intervalo 1*	*Intervalo 2*	*Intervalo 3*
0,10	0,1054	0,19	0,1476	(0,0979; 0,1973)	(0,1089; 0,2095)	(0,0878; 0,2073)
0,25	0,2877	0,30	0,3480	(0,2647; 0,4313)	(0,2775; 0,4463)	(0,2461; 0,4499)
0,50	0,6931	0,30	0,7482	(0,6129; 0,8835)	(0,6211; 0,9026)	(0,6001; 0,8963)
0,75	1,3863	0,51	1,2363	(0,9928; 1,4797)	(1,0161; 1,4949)	(1,1873; 1,2851)
0,90	2,3026	0,10	2,5712	(1,9456; 3,1969)	(1,8124; 3,0535)	(2,1278; 3,0146)
0,95	2,9957	0,08	2,7972	(2,2640; 3,3304)	(2,2905; 3,2648)	

próximos dos respectivos valores teóricos $Q^0(p)$. Em compensação os intervalos da Tabela 3.12 apresentam uma amplitude menor do que na Tabela 3.11. O que é comum entre as duas tabelas é a assimetria dos intervalos. Não foi possível calcular o *Intervalo 3* da Tabela 3.12 para o percentil $0,95$.

Capítulo 4

Aplicações

4.1 Exemplo da Aplicacão 1

O objetivo é estimar a distribuição do tempo de vida de 40 interruptores mecânicos, dos quais $57,5\%$ foram censurados. Mais detalhes podem ser vistos em Nair[1984]. O conjunto de dados encontra-se no Apêndice B.

A Tabela 4.1 apresenta um comparativo entre as janelas estimadas em Padgett[1986] e as estimadas neste trabalho. Seguindo a mesma metodologia no artigo referido trocando

$$Q_n(p) = \sum \frac{1}{h} \int_{S_{i-1}}^{S_i} K\left(\frac{t-p}{h}\right)$$

pela mesma equação, mas com o cálculo da integral realizado analiticamente.

As janelas selecionadas em Padgett[1986] são crescentes com aumento do percentil indicando uma maior suavidade para percentis próximos a 1. Já as janelas encontradas neste texto oscilam, sendo que a partir da mediana elas decrescem.

A estimativa para a mediana calculada em Pagett[1986] foi de $Q_n^*(0,5) = 2,587$, a estimativa usando o estimador de Kaplan-Meier foi de $\tilde{Q}_n(0,5) = 2,548$, e o estimador encontrado neste trabalho foi $Q_n(0,5) = 2,525$, observando que esta última

Tabela 4.1: Estimativas das janelas

Quantil	0,1	0,25	0,5	0,75	0,90	0,95
h_n^*	0,30	0,26	0,34	0,34	0,40	0,47
h_n	0,17	0,05	0,59	0,30	0,18	0,01

Tabela 4.2: Estimativas Intervalares para Mediana

$Intervalo1$	(2,3063; 2,7431)
$Intervalo2$	(2,3425; 2,7751)
$Intervalo3$	(2,3538; 2,6956)

Tabela 4.3: Estimativas dos percentis

Quantil	0,1	0,25	0,5	0,75	0,90	0,95
$\tilde{Q}_n$	1,710	2,197	2,548	3,015	3,017	3,017
Q_n	1,665	2,179	2,525	2,989	3,019	3,793

estimativa está próximo da estimativa de Kaplan-Meier. Também foram calculados os três intervalos de confiança para a mediana (Tabela 4.2).

Na Tabela 4.3, estão estimativas de Kaplan-Meier e as estimativas calculadas nesta dissertação usando bootstrap através do núcleo, para alguns percentis.

4.2 Exemplo da Aplicacão 2

O conjunto de dados é referente ao estudo de crianças com leucemia, desenvolvido pelo Grupo Mineiro para Tratamento de Leucemias Agudas. Um grupo de 128 crianças, com idade inferior a 15 anos, foi acompanhado no período de 1988 a 1992, em alguns hospitais de Belo Horizonte. A variável de interesse é o tempo a partir da remissão da doença (ausência da doença) até a reciditiva ou morte (o que ocorrer primeiro). Das 128 crianças, 120 entraram em remissão, e este será o grupo de estudo. Os dados estão no Apêndice B e mais detalhes pode ser encontrado em Colosimo[2002].

As estimativas para os percentis 0,1 e 0,25, usando o estimador de Kaplan-Meier, foram de $\tilde{Q}_n(0,1) = 0,572$ e $\tilde{Q}_n(0,25) = 1,306$, já para o estimador usando o núcleo com amostras bootstrap, foram de $Q_n(0,1) = 0,569$ e $Q_n(0,25) = 1,402$.

Devido a alta proporção de censura(72%), não foi possível estimar os percentis a partir da mediana.

Capítulo 5

Conclusão

Neste trabalho, foi proposto avaliar as estimativas das funções de densidade e quantílica usando o método do núcleo estimador.

Para a função de densidade, a janela selecionada subestimou a janela téorica, usando o método de Validação Cruzada, quando foram realizadas simulações com amostras exponenciais, vindo ao encontro dos resultados de Chagas[2004] para amostras normais.

Na estimação pontual para a função quantílica, as simulações usando distribuições Weibull e Lognormal com vários tamanhos de amostras com o método do Bootstrap, apresentaram uma grande variabilidade na seleção das janelas.

Acredita-se que a origem desta variabilidade esteja na geração das amostras, devido às amostras não bem ajustadas. Tomando a média destas janelas, as estimativas dos quantis apresentaram-se próximas do valor teórico. Mesmo aquelas janelas selecionadas com baixa frequência e distantes da média resultaram numa boa estimação para o quantil.

Também foram calculadas aproximações para intervalos de confiança. Utilizou-se três métodos, sendo que o método proposto em Cheng[2002] aprensentou melhores resultados para percentis abaixo da mediana. Para os percentis 0,75, 0,90 e 0,95, nenhum dos três métodos teve bom desempenho.

Para propostas futuras, sugere-se estimar a janela para cada percentil e assim construir a função quantil e estimar a função de densidade usando o bootstrap.

Apêndice A

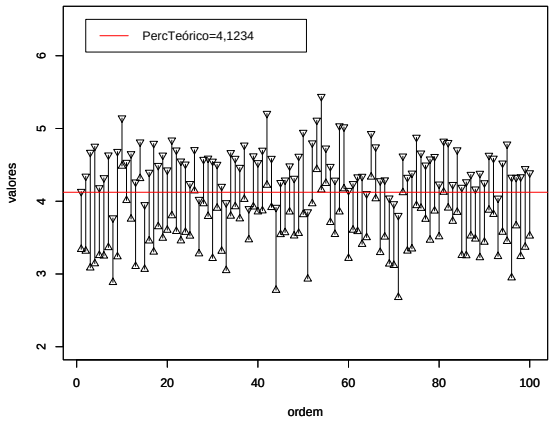

Figura 5.1: Int1; Q=0,1; n=50; WB

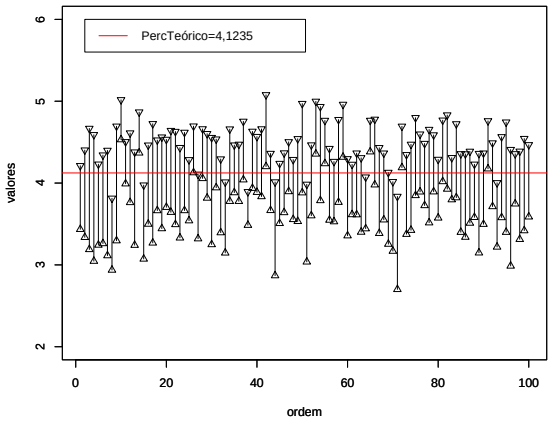

Figura 5.2: Int2; Q=0,1; n=50; WB

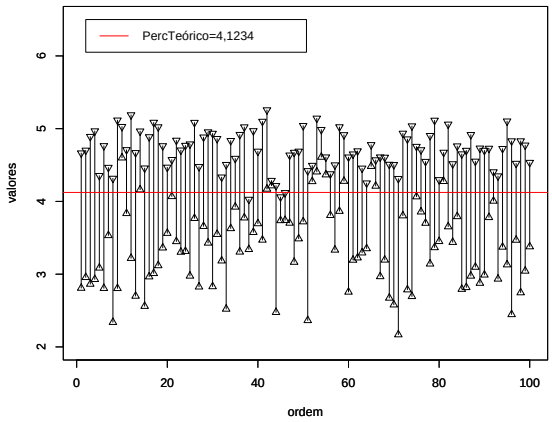

Figura 5.3: Int3; Q=0,1; n=50; WB

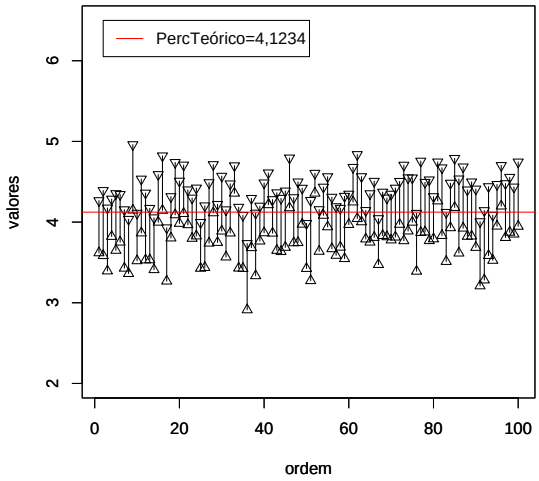

Figura 5.4: Int1; Q=0,1; n=100; WB

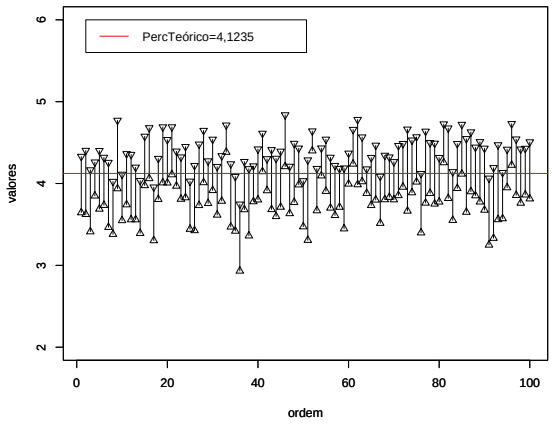

Figura 5.5: Int2; Q=0,1; n=100; WB

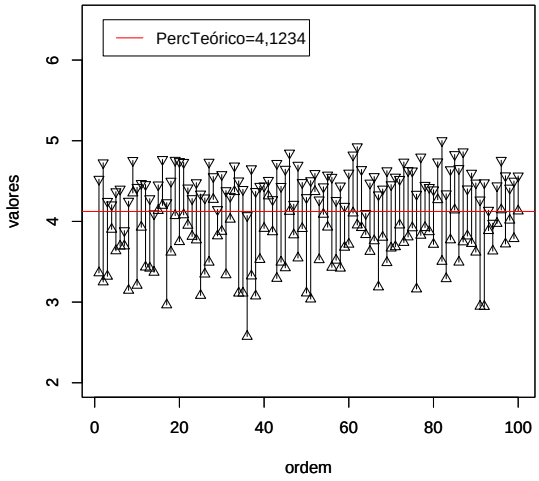

Figura 5.6: Int3; Q=0,1; n=100; WB

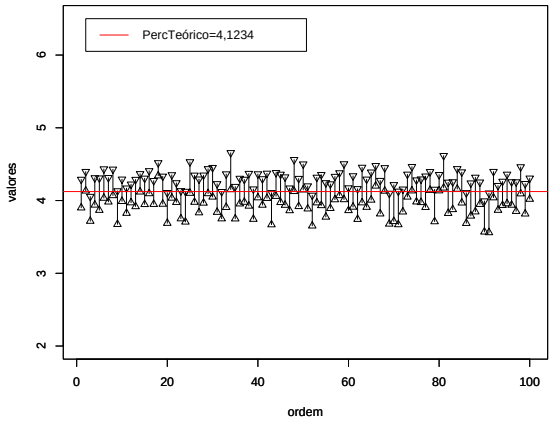

Figura 5.7: Int1; Q=0,1; n=300; WB

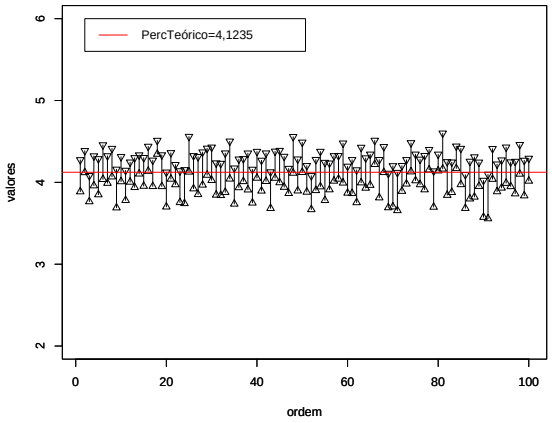

Figura 5.8: Int2; Q=0,1; n=300; WB

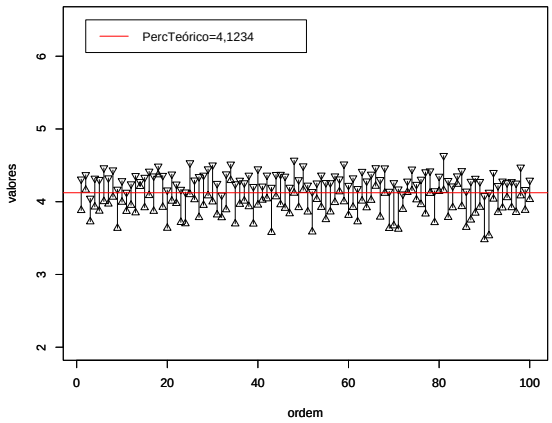

Figura 5.9: Int3; Q=0,1; n=300; WB

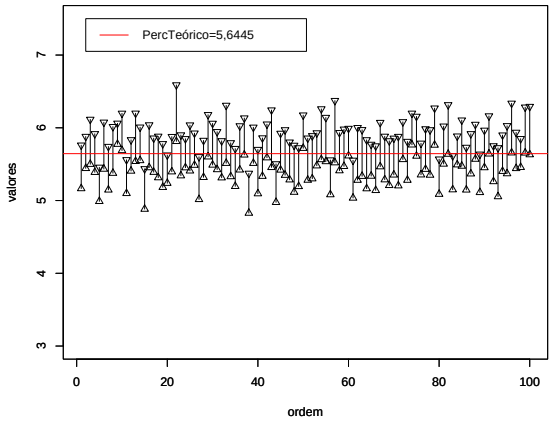

Figura 5.10: Int1; Q=0,5; n=50; WB

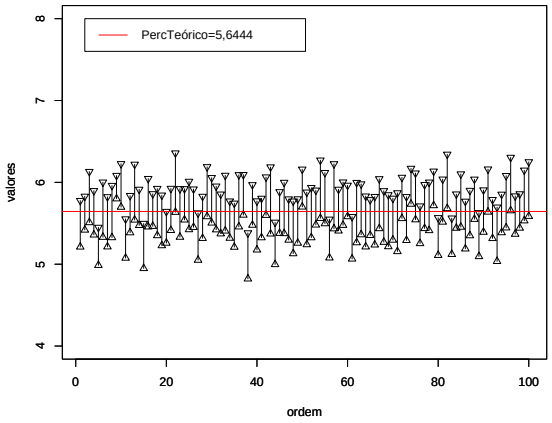

Figura 5.11: Int2; Q=0,5; n=50; WB

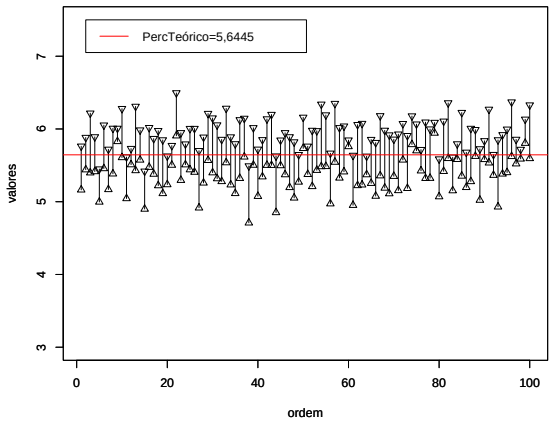

Figura 5.12: Int3; Q=0,5; n=50; WB

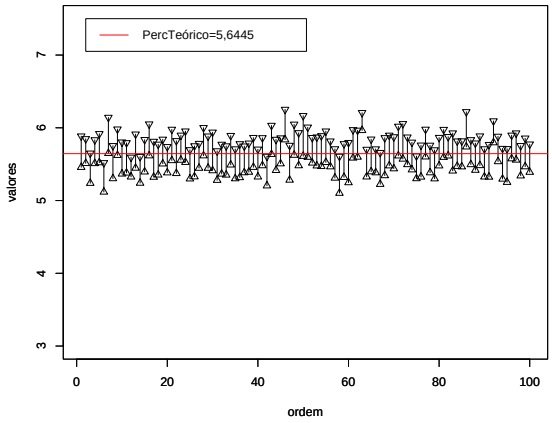

Figura 5.13: Int1; Q=0,5; n=100; WB

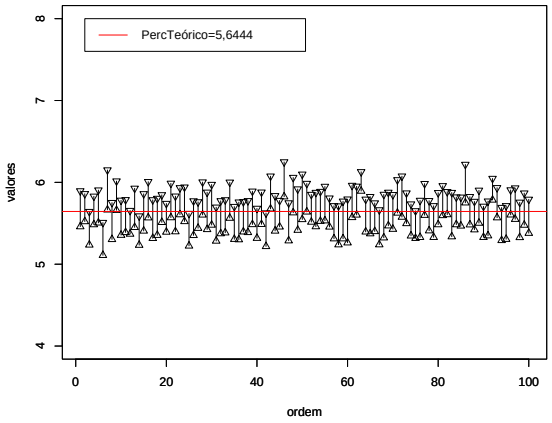

Figura 5.14: Int2; Q=0,5; n=100; WB

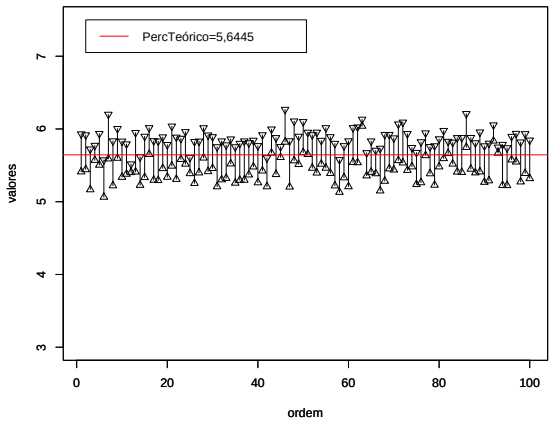

Figura 5.15: Int3; Q=0,5; n=100; WB

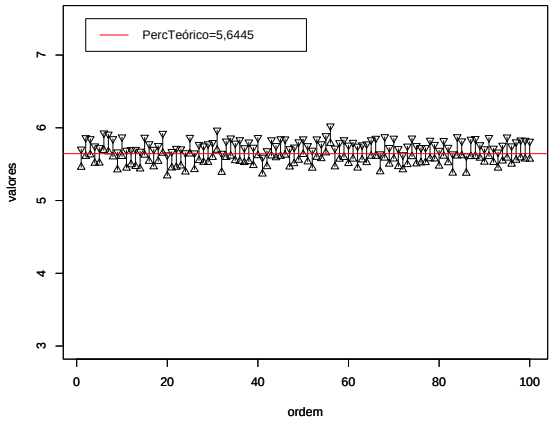

Figura 5.16: Int1; Q=0,5; n=300; WB

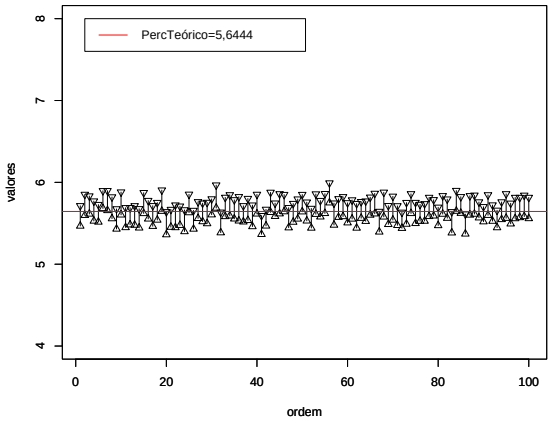

Figura 5.17: Int2; Q=0,5; n=300; WB

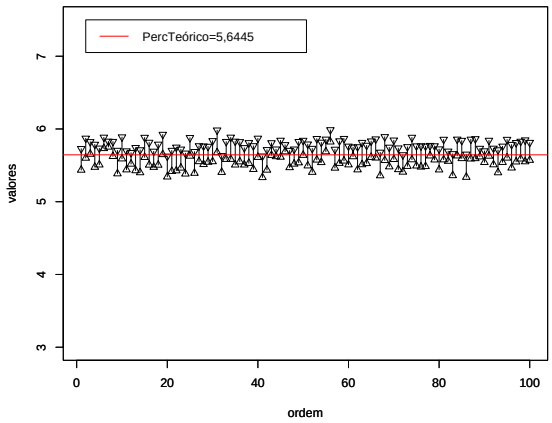

Figura 5.18: Int3; Q=0,5; n=300; WB

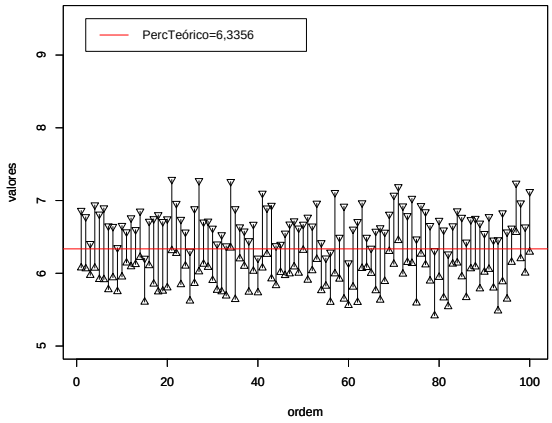

Figura 5.19: Int1; Q=0,75; n=30; WB

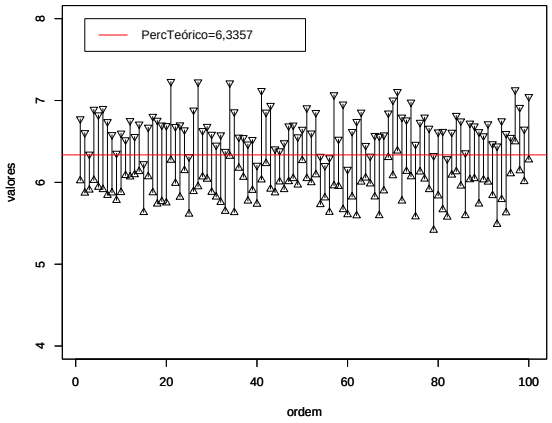

Figura 5.20: Int2; Q=0,75; n=30; WB

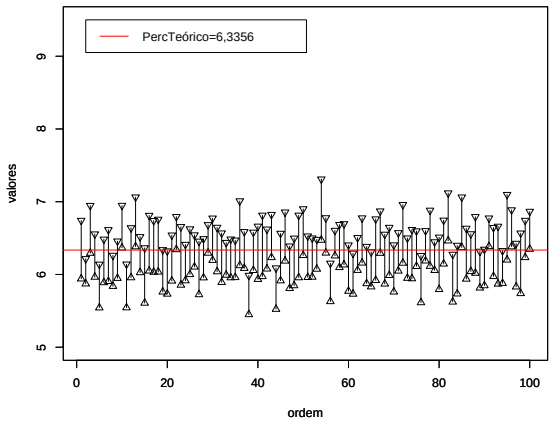

Figura 5.21: Int1; Q=0,75; n=50; WB

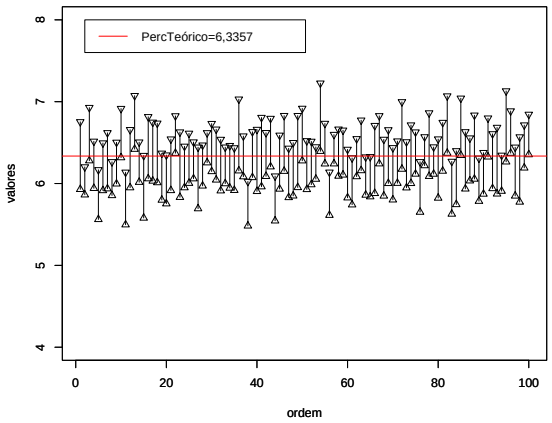

Figura 5.22: Int2; Q=0,75; n=50; WB

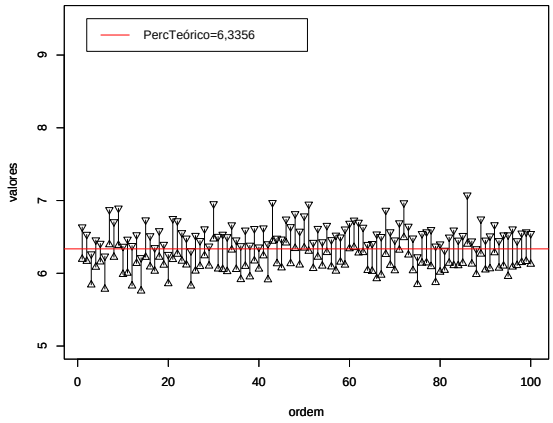

Figura 5.23: Int1; Q=0,75; n=100; WB

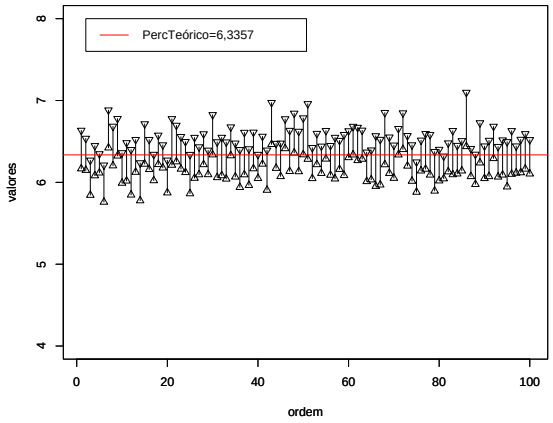

Figura 5.24: Int2; Q=0,75; n=100; WB

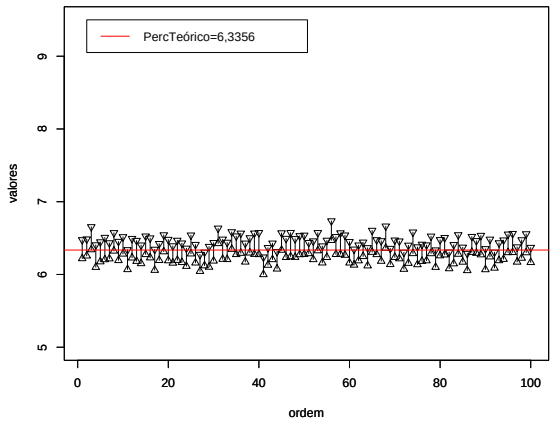

Figura 5.25: Int1; Q=0,75; n=300; WB

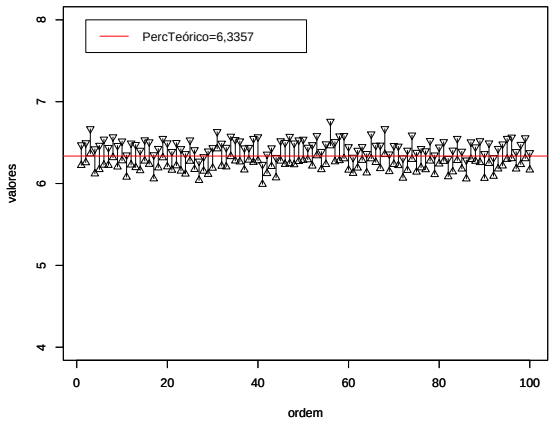

Figura 5.26: Int2; Q=0,75; n=300; WB

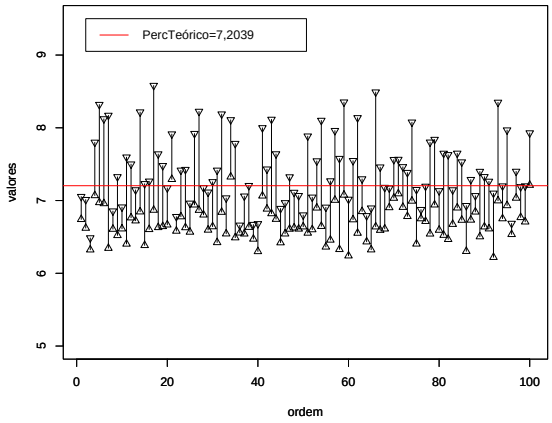

Figura 5.27: Int1; Q=0,95; n=30; WB

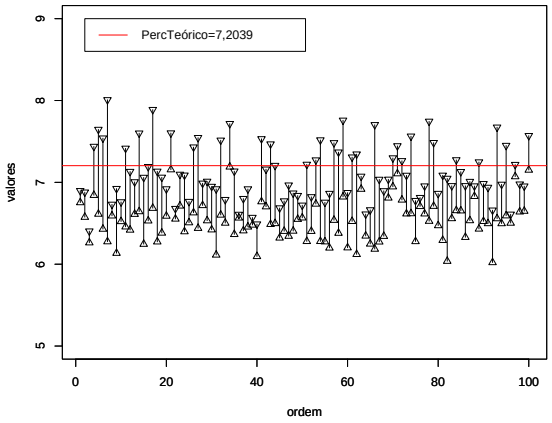

Figura 5.28: Int2; Q=0,95; n=30; WB

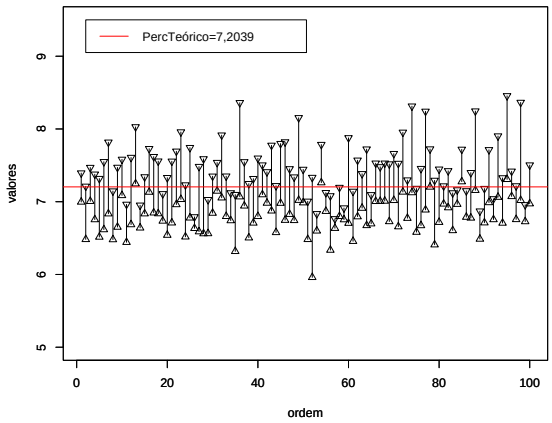

Figura 5.29: Int1; Q=0,95; n=50; WB

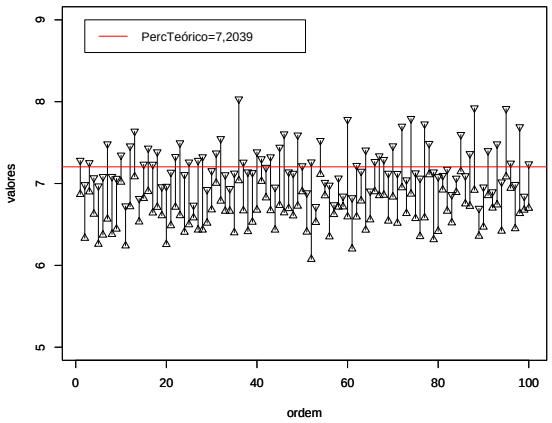

Figura 5.30: Int2; Q=0,95; n=50; WB

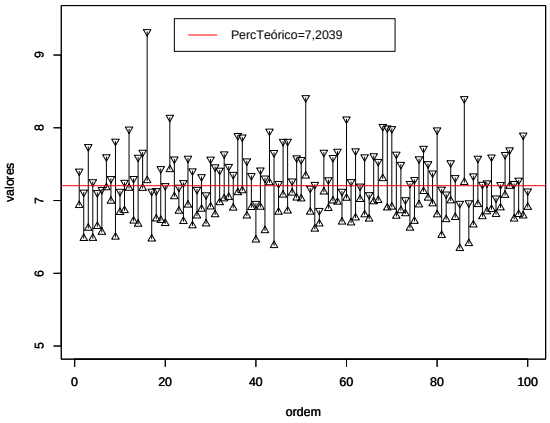

Figura 5.31: Int1; Q=0,95; n=100; WB

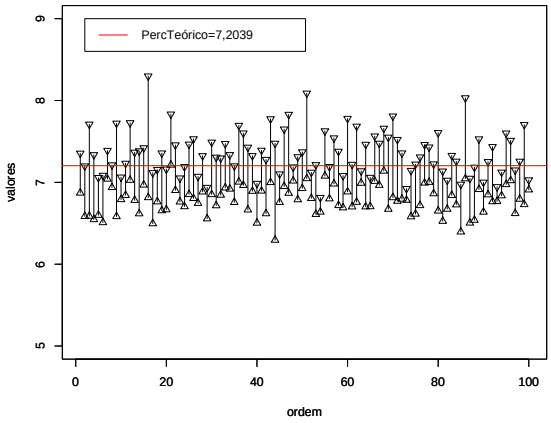

Figura 5.32: Int2; Q=0,95; n=100; WB

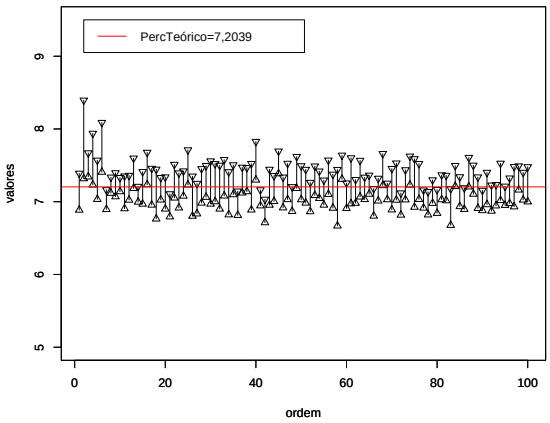

Figura 5.33: Int1; Q=0,95; n=300; WB

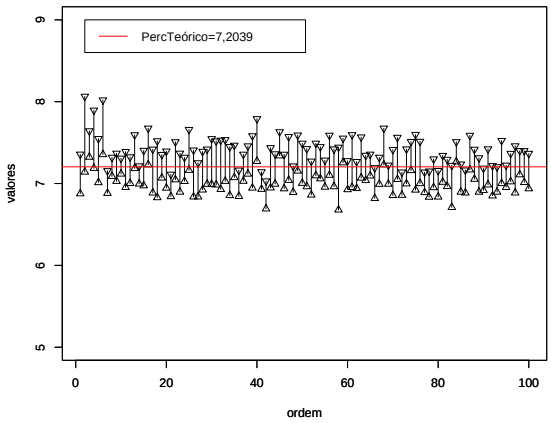

Figura 5.34: Int2; Q=0,95; n=300; WB

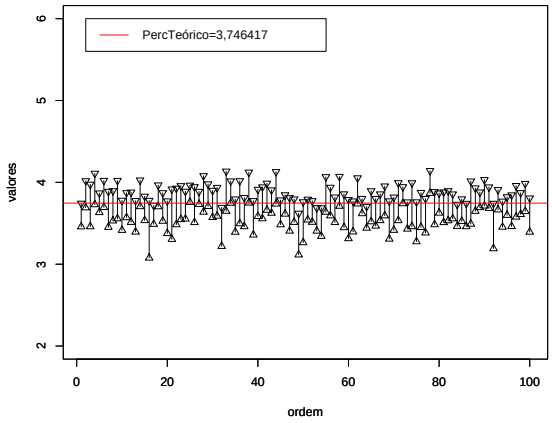

Figura 5.35: Int1; Q=0,1; n=50; LN

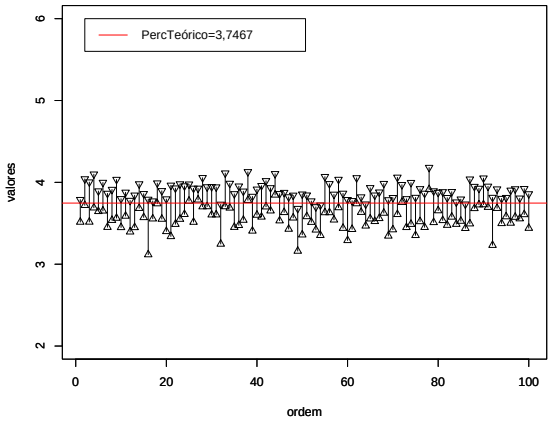

Figura 5.36: Int2; Q=0,1; n=50; LN

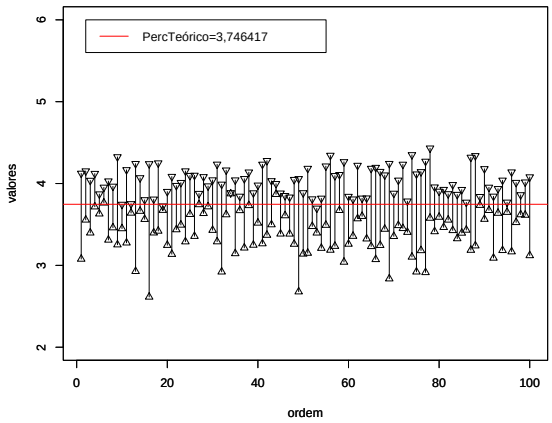

Figura 5.37: Int3; Q=0,1; n=50; LN

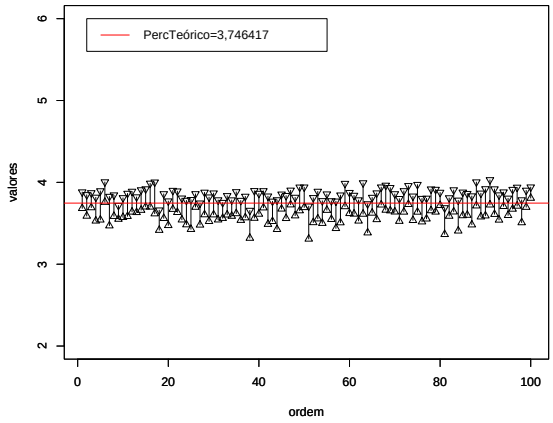

Figura 5.38: Int1; Q=0,1; n=100; LN

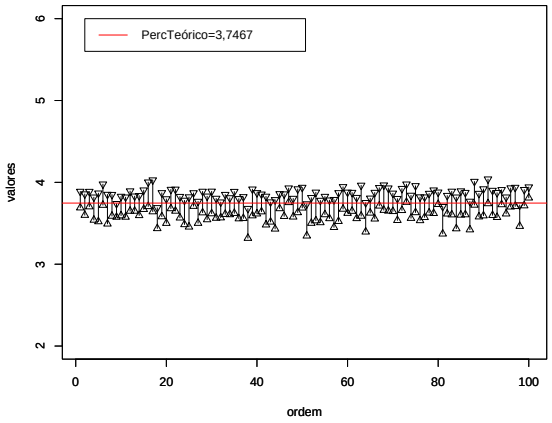

Figura 5.39: Int2; Q=0,1; n=100; LN

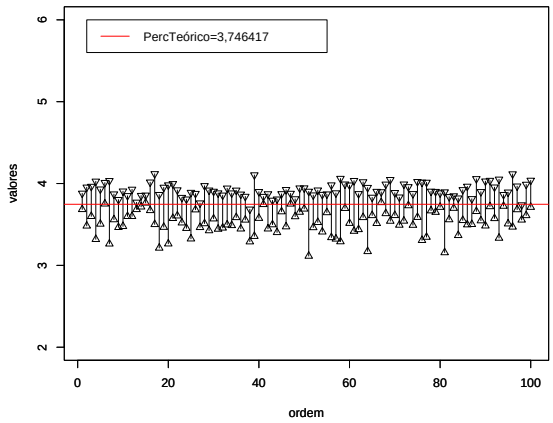

Figura 5.40: Int3; Q=0,1; n=100; LN

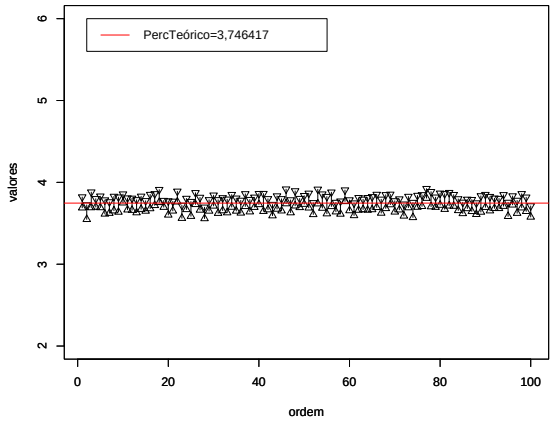

Figura 5.41: Int1; Q=0,1; n=300; LN

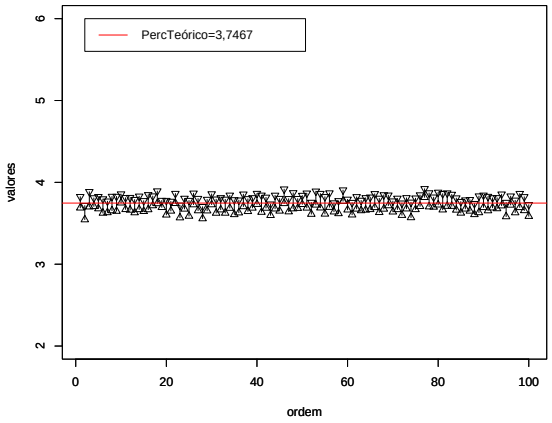

Figura 5.42: Int2; Q=0,1; n=300; LN

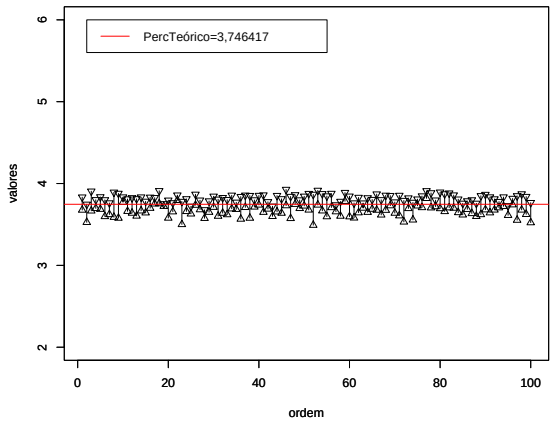

Figura 5.43: Int3; Q=0,1; n=300; LN

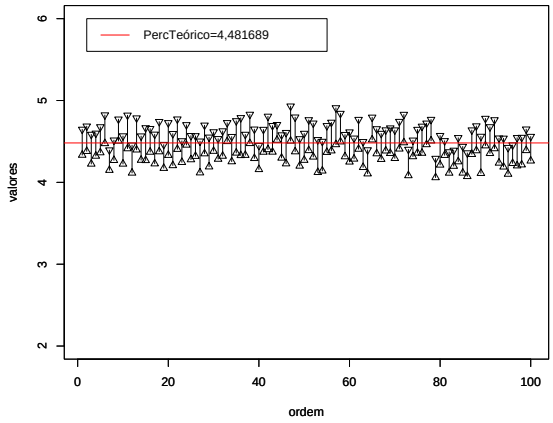

Figura 5.44: Int1; Q=0,5; n=50; LN

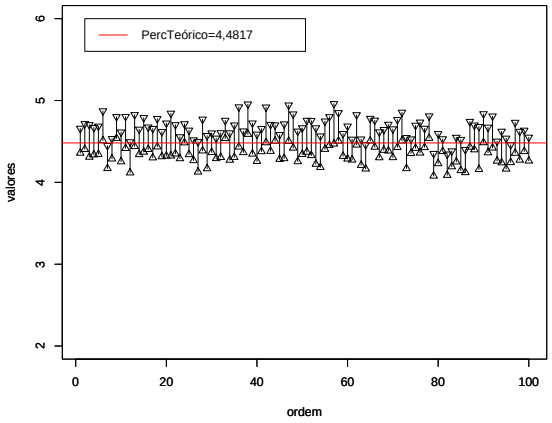

Figura 5.45: Int2; Q=0,5; n=50; LN

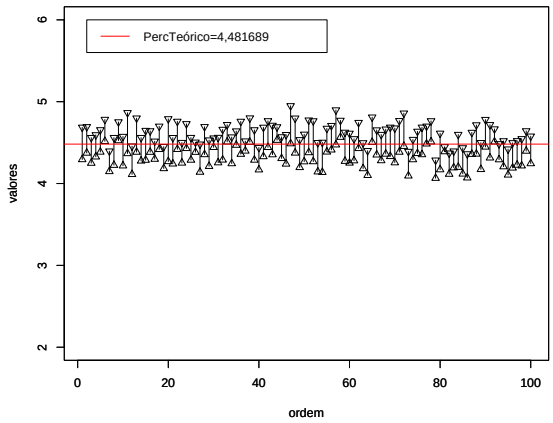

Figura 5.46: Int3; Q=0,5; n=50; LN

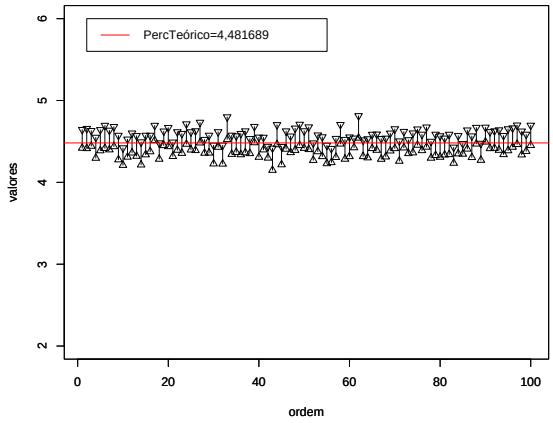

Figura 5.47: Int1; Q=0,5; n=100; LN

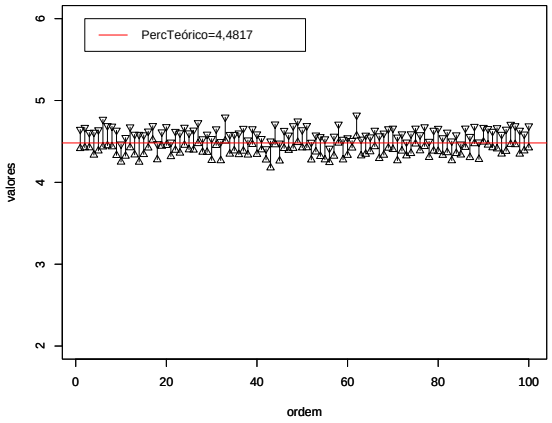

Figura 5.48: Int2; Q=0,5; n=100; LN

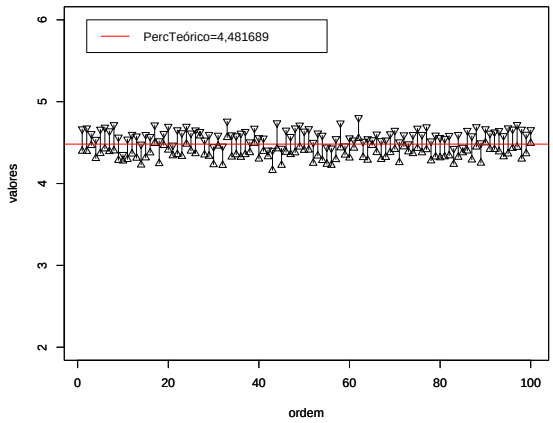

Figura 5.49: Int3; Q=0,5; n=100; LN

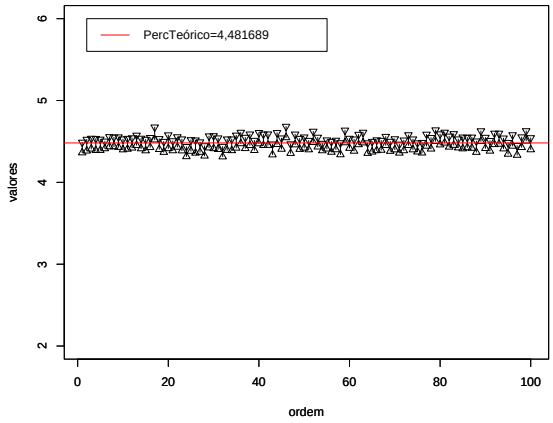

Figura 5.50: Int1; Q=0,5; n=300; LN

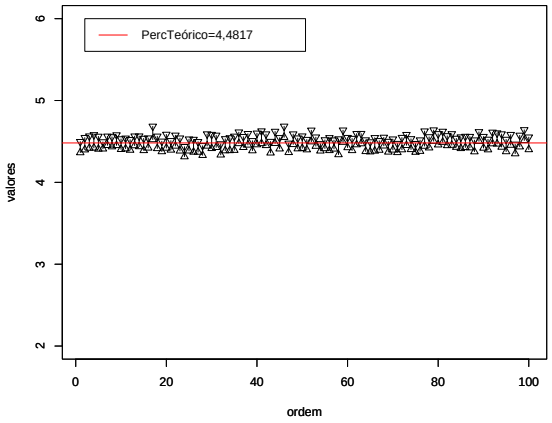

Figura 5.51: Int2; Q=0,5; n=300; LN

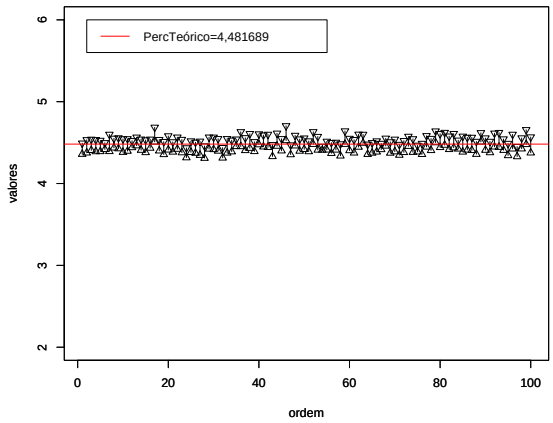

Figura 5.52: Int3; Q=0,5; n=300; LN

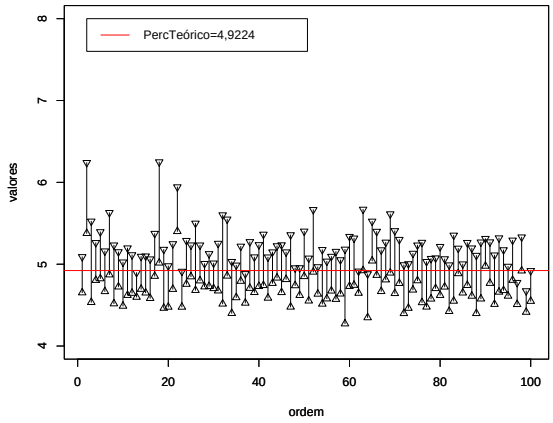

Figura 5.53: Int1; Q=0,75; n=30; LN

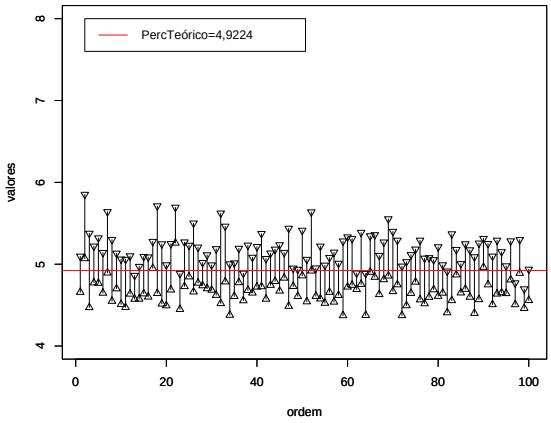

Figura 5.54: Int2; Q=0,75; n=30; LN

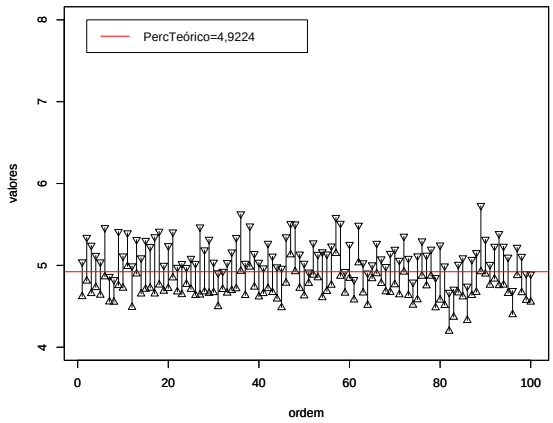

Figura 5.55: Int1; Q=0,75; n=50; LN

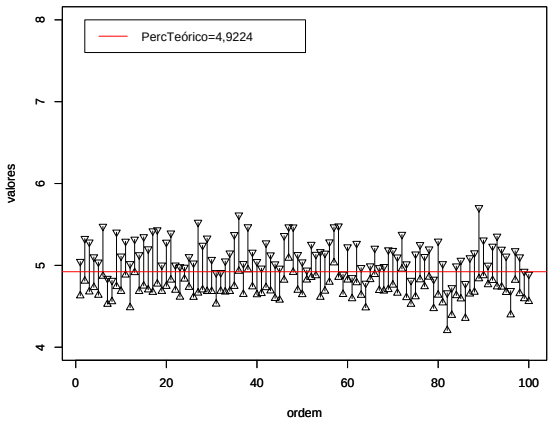

Figura 5.56: Int2; Q=0,75; n=50; LN

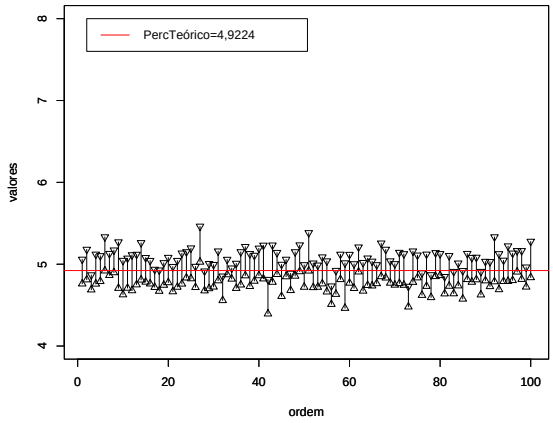

Figura 5.57: Int1; Q=0,75; n=100; LN

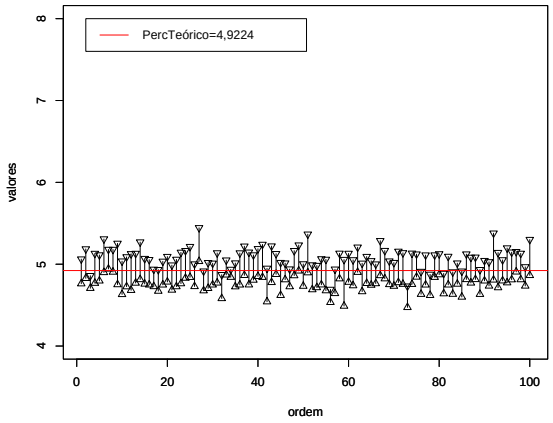

Figura 5.58: Int2; Q=0,75; n=100; LN

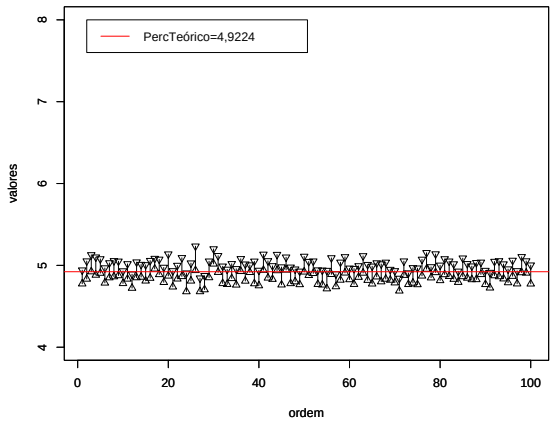

Figura 5.59: Int1; Q=0,75; n=300; LN

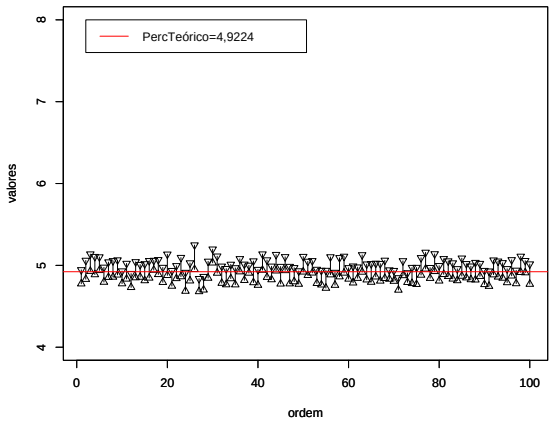

Figura 5.60: Int2; Q=0,75; n=300; LN

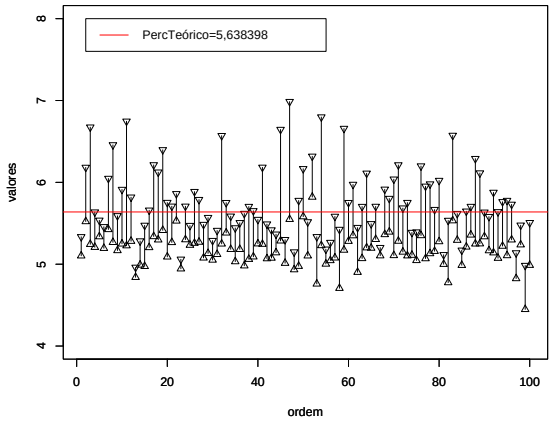

Figura 5.61: Int1; Q=0,95; n=30; LN

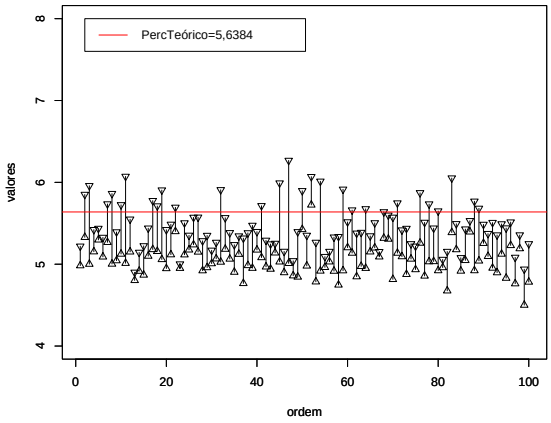

Figura 5.62: Int2; Q=0,95; n=30; LN

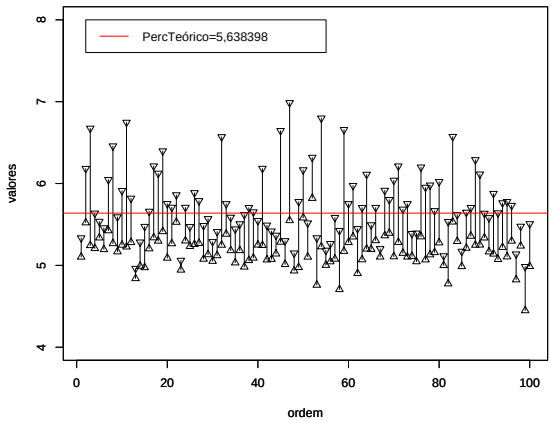

Figura 5.63: Int1; Q=0,95; n=50; LN

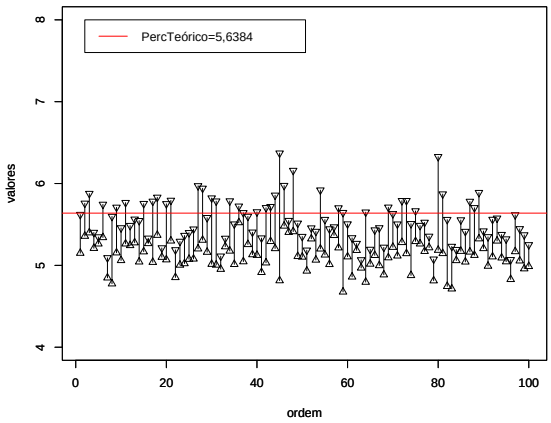

Figura 5.64: Int2; Q=0,95; n=50; LN

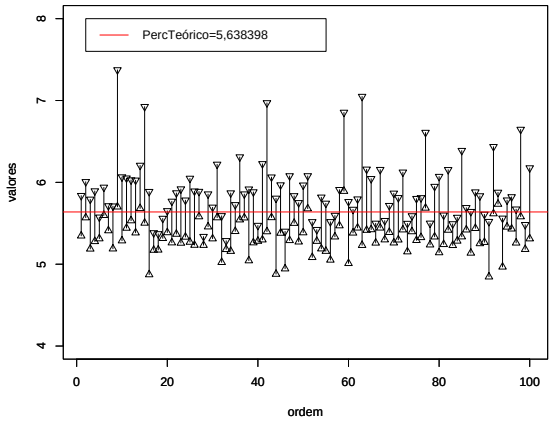

Figura 5.65: Int1; Q=0,95; n=100; LN

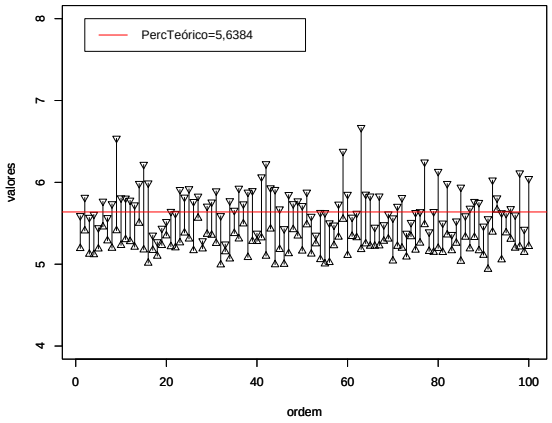

Figura 5.66: Int2; Q=0,95; n=100; LN

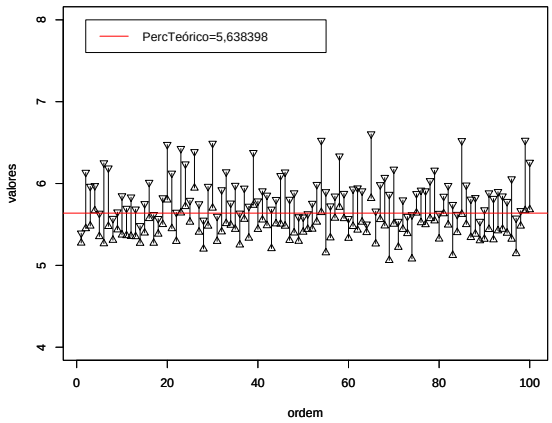

Figura 5.67: Int1; Q=0,95; n=300; LN

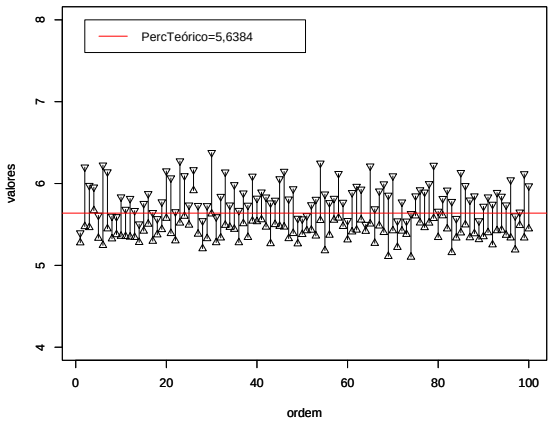

Figura 5.68: Int2; Q=0,95; n=300; LN

Apêndice B

Tabela 5.1: Conjunto de Dados dos Interruptores

T	δ	T	δ	T	δ	T	δ
1,151	0	1,667	1	2,119	0	2,547	1
1,17	0	1,695	1	2,135	1	2,548	1
1,248	0	1,71	1	2,197	1	2,788	0
1,331	0	1,955	0	2,199	0	2,794	1
1,381	0	1,965	1	2,227	1	2,883	0
1,499	1	2,012	0	2,25	0	2,883	0
1,508	0	2,051	0	2,254	1	2,91	1
1,543	0	2,076	0	2,261	0	3,015	1
1,577	0	2,109	1	2,349	0	3,017	1
1,584	0	2,116	0	2,369	1	3,793	0

Tamanho da amostra n=40;

$\delta = 1$ indica tempo de falha;

Detalhes Nair[1984];

Tabela 5.2: Conujunto de Dados de Leucemia

T	δ	T	δ	T	δ	T	δ
0,003	0	0,81	0	1,566	1	2,738	0
0,003	0	0,841	1	1,624	0	2,757	0
0,025	1	0,843	1	1,692	0	2,762	1
0,099	1	0,893	0	1,697	0	2,765	0
0,129	1	0,895	0	1,714	0	2,861	0
0,151	0	0,898	0	1,76	1	2,902	1
0,214	1	0,917	1	1,832	0	3,006	0
0,26	1	0,969	0	1,927	0	3,014	0
0,268	0	0,991	0	1,941	0	3,11	0
0,383	0	0,991	0	1,96	0	3,198	0
0,454	1	0,994	0	2,053	1	3,209	0
0,46	1	1,065	0	2,07	1	3,403	0
0,487	1	1,103	0	2,119	1	3,425	0
0,498	1	1,12	0	2,133	0	3,466	1
0,572	1	1,18	0	2,171	0	3,485	0
0,594	1	1,205	0	2,22	0	3,518	0
0,616	1	1,227	1	2,264	1	3,578	0
0,63	1	1,259	0	2,278	0	3,639	1
0,654	1	1,27	1	2,333	1	3,704	0
0,687	1	1,306	1	2,344	0	3,754	0
0,701	0	1,322	1	2,355	0	3,83	0
0,709	1	1,41	0	2,428	0	3,896	0
0,715	1	1,443	1	2,464	0	3,915	0
0,736	1	1,481	0	2,502	1	4,252	0
0,742	0	1,52	0	2,639	0	4,331	0
0,758	0	1,528	0	2,65	1		

Tamanho da amostra n=103;

$\delta = 1$ indica tempo de falha;

Apêndice C

Início Algoritmo

Gerar dados

Executar cálculos para a Amostra Original

Executar cálculos para as Amostras Bootstrap

Calcular o EQM entre as estimativas da Amostra Original com das Amostras Bootstrap

Calcular a Média das janelas selecionada

Recalcular Qn com a média do passo anterior

Fazer a escolha da Janela

Calcular o Intervalo de Confiança de Qn usando cada janela selecionada

Saída de resultados

Fim Algoritmo

Ref. Gerar dados

Gerar duas amostras de tamanho ajustáveis e alocá-las nos vetores Tempo_ de_ Vida e Tempo_ de_ Censura

Tomar o mínimo entre cada par gerado e alocá-las em Valor_ Gerado

Se $Valor\_Gerado[i] == Tempo\_de\_Vida[i]$ *atribuir a Indicador[i] o valor 1*

Senão atribuir a Indicador[i] o valor 0

Ordenar Valor_ Gerado pareado a Indicador

Gerar matriz de inteiros de tamanho $(300xn)$ *que gerará a matriz de Bootstrap*

Fim Ref.

Ref. Cálculos para a Amostra Original

Calcular a função H para o tempo de sobrevivência*

Calcular a função Pn(t)

Calcular a função salto sj

Calcular a função Qn

Calcular o estimador de Kaplan-Meyer da função quantil

Fim Ref.

Ref. Cálculos para as Amostras Bootstrap

Calcular a função H para o tempo de sobrevivência*

Calcular a função Pn(t)

Calcular a função salto sj

Calcular a função Qn

Calcular o estimador de Kaplan-Meyer da função quantil

Fim Ref.

Referências Bibliográficas

[1] Aly, E. E., Csörgo, M., e Horváth, L., 1985. "Strong Approximations of the Quantile Process of the Product-Limit Estimator". *Journal of Multivariate Analysis.* **16** 185-210.

[2] Ayer, M., Brunk, H. D., Ewing, G. M, Reid, W. T., Silverman, E., 1955. "An Empirical Distribution Function for Sampling with Incomplete Information". *Ann. Math. Statist.* **26**, 641-647.

[3] Bessegato, L. F., 2001. "Escolha do Parâmetro de Suavidade na Estimativa da Função de Distribuição". *Dissertação de Mestrado, Departamento de Estatística da UFMG.*

[4] Biao Zhang, 1996. "Some asymptotic results for kernel density estimation under random censorship". *Bernoulli, Chapman Hall, Department of Mathematics, The University of Toledo, Toledo OH 43606, USA*, 183198.

[5] Bowman, A., 1984 . "An Altenative Method of Cross-Validation for the Smoothing of Density Estimates". *Biometrika.***71** , 353-360.

[6] Blum, J. R. e Sursala, V.,1980. "Maximal Deviation Theory of Density and Failure Estimates Based on Censored Data". *In Multivariate Analysis V (***P. R. Krishnaiah***)*, 213-222.

[7] Breslow, N. E. e Crowley, J., 1974. " A Large Sample Study of the Life Table and Product Limit Estimates Under Random Censorship". *The Annals of Statistics,* **2**, 437-53.

[8] Cao, R. and Jácome, M. A., 2003. "Presmoothed kernel density estimator for

censored data." *Technical Report. Reports in Statistics and Operations Research, University of Santiago de Compostela.*

[9] Chagas, R. M., 2004. "Avaliação do Método de Validação Cruzada para Estimar a janela Ótima-Dados Censurados". *Relatórios de Projetos em Estatística, Departamento de Estatística-UFMG.*

[10] Cheng, C., 2002. "Almost Sure Uniform Error Bounds of General Smooth Estimators of Quantile Density Function". *Statistics and Probability Letters,* **59**, 183-194.

[11] Cheng, K. F., 1984. "On Almost Sure Representations for Quantiles of the Product Limit Estimator with Applications". *Sankhya, Ser. A,* **46**, 426-443.

[12] Chiu, S. T., 1991. " Bandwidth Seletion for Kernel Density Estimation". *The Annals of Statistics,* **19**, 1883-1905.

[13] Colosimo, E. A., 2001."Análise de Sobrevivência Aplicada". *46ª Reunião Anual da RBRAS e 9º SEAGRO. Departamento de Estatística da UFMG.*

[14] Csörgo, M., 1983. "Quantile Processes with Statistical Applications". *CBMS-NSF Regional Conference Series in Applied Mathematics*, Philadelphia: Society for Industrial and Applied Mathematics.

[15] Efron, B., 1979. " Bootstrap methods: Another Look at the Jackknife". *Ann. Statist.* **7**, 1-26.

[16] Efron, B., 1967. "The Two-Sample Problem With Censored Data". *In Proceedings of the Fifth Berkeley Symposium***Vol. 4**,Ca:University of California Press, 831-853.

[17] Efron, B., Tibshirani, R., 1993. "An Introduction to the Bootstrap". *Chapman & Hall, New York.*

[18] Hall, P., 1986. "On the Number of Bootstrap Simulations Required to Construct a Confidence Interval". *Ann. Statist.,* **14**, 1453-1462.

[19] Kaplan, E. L., e Meier, P., 1958. "Nonparametric Estimation from Incomplete Observations,". *Journal of the American Statistical Association.,* **53**, 457-481.

[20] Lio, Y. L., Padgett, W.J., and Yu, K., F., 1986. "On the Asymptotic Properties of a Kernel-Type Quantile Estimator From Censored Samples". *Statistics Technical Reports 104, University of South Carolina.*

[21] Mathews, J. H., 1987. "For Computer Science, Engineering, And Mathematics". *Prentice-Hall, Inc.*

[22] Marron, J. S.e Padgett, W. J., 1987. "Asymptotically Optimal Bandwidth Selection for Kernel Density Estimators from Randomly Right-Censored Samples". *The Annals of Statistics,* **4**, 1520-1535.

[23] Nair, V. N., 1984. "Confidence Bands for Survival Functions with Censored Data: A Comparative Study,". *Technometrics,* **26**, 265-275.

[24] Padgett, W. J., 1986. "A Kernel-type Estimator of a Quantile Function from Right-Censored Data". *Journal of the American Statistical Association,* **39**, 215-222.

[25] Padgett, W. J. e McNichols, D. T, 1984."Nonparametric Density Estimation From Censored Data". *Communications in Statistics-Theory and Methods*,**13** 1581-1611.

[26] Padgett, W. J. e Thombs, L. A., 1986. "Smooth Nonparametric Quantile Estimators Under Censoring: Simulations and Bootstrap Methods". *Comun. Statistics-Simula,* **15**, 1003-1025.

[27] Parzen, E., 1962. " On the Estimation of a Probability Density Function and the Mode". *Ann. Math. Statist.* **33**, 1065-76.

[28] Rosenblatt, M., 1956. " Remarks on Some Nonparametric Estimates of a Density Function". *Ann. Math. Statist.* **27**, 832-7.

[29] Rosenblatt, M., 1976. "On the Maximal Deviation of k-Dimensional Density Estimates". *Ann. Probab.* **4**, 1009-1015.

[30] Rudemo, M. 1982., "Empirical Choice of Histograms and Kernel Density Estimators". *Scand J. Statist.* **9**, 65-78.

[31] Sander, J. 1975., "The Weak Convergence of Quantiles of the Product Limit estimator". *Technical Report 5, Stanford University, Dept. of Statistics.*

[32] Silverman, B. W., 1986. " Density Estimation for Statistics and Data Analysis". *Chapman & Hall, London.*

[33] Simonoff, J. S., 1996. " Smoothing Methods in Statistics". *Springer Series in Statistics.*

[34] Yang, B., 1985. "A Smooth Nonparametric Estimator of a Quantile Function". *Journal of American Statistical Association.***80**, 1004-1011.