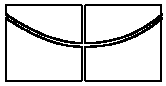


Universidade de Brasília

Instituto de Ciências Exatas

Departamento de Estatística

Dissertação de Mestrado

Estimação da função quantílica para dados com censura intervalar

por

André Luís Silva

Orientador: Prof. Dr. Antonio Eduardo Gomes

Junho de 2010

André Luís Silva

Estimação da função quantílica para dados com censura intervalar

Dissertação apresentada ao Departamento de Estatística do Instituto de Ciências Exatas da Universidadede de Brasília como requisito parcial à obtenção do título de Mestre em Estatística.

Universidade de Brasília

Brasília, Junho de 2010

A Paulo Sergio Silva (in memorian) e a
Maria Aparecida da Cunha Silva, meus pais,
por tudo o que fizeram por mim e pelo que
representam em minha vida.

Agradecimentos

- Agradeço inicialmente a Deus, por que dele, por ele e para ele são todas as coisas.
- A Tatiane Silva, minha querida esposa, pelo amor, pelo carinho, pela dedicação e, principalmente, por me compreender nos momentos em que estive ausente.
- Ao professor Antonio Eduardo, pela orientação e amizade.
- A todos aqueles que contribuíram direta ou indiretamente para a realização deste trabalho.

Sumário

Lista de Figuras

Lista de Tabelas

Resumo

Este trabalho propõe estimar a função quantílica na presença de censura intervalar, com especial atenção aos quantis mais baixos. Para tanto, inicialmente adota-se a abordagem não-paramétrica para obter o estimador de máxima verossimilhança da função de sobrevivência, mediante o emprego da teoria da regressão isotônica. Utiliza-se o Núcleo Estimador para suavizar a estimativa de máxima verossimilhança, por intermédio de vários métodos de obtenção do parâmetro ótimo de suavização. Por fim, estima-se a função quantílica e compara-se o vício e a variabilidade das estimativas para diferentes tamanhos de amostra e padrões de intervalos de censura.

Palavras Chave: *Estimador não Paramétrico de Máxima Verossimilhança, Regressão Isotônica, Núcleo Estimadores, Censura Intervalar, Função Quantílica.*

Abstract

In this work, we study the estimation of the quantile function, especially for small quantiles. For that, we used the nonparametric maximum likelihood estimator (NPMLE) of the survival function, which is obtained using the theory of isotonic regression. We used kernel smoothing for the NPMLE with different methods to obtain the bandwidth. We compared the bias and variance of the estimators for different sample sizes and interval censoring patterns.

key words: *Nonparametric Maximum Likelihood Estimator, Isotonic Regression, Kernel Smoothing, Interval Censoring, Quantile Function.*

Capítulo 1

Introdução

1.1 Motivação

Dados com censura intervalar ocorrem quando sabe-se apenas que o valor da variável de interesse, T, geralmente definida como o tempo até a ocorrência de um evento de interesse ("falha"), pertence a um intervalo (U, V).

A variável T, comumente denominada tempo de falha, pode representar o tempo até a morte de um paciente, ou até a cura ou recidiva de uma doença, ou o tempo até a falha de uma munição. Pode ser expressa em outras unidades de medida, por exemplo, o número de quilômetros rodados até que um pneu esteja desgastado.

Em muitos casos, deseja-se estimar os quantis da variável T como, por exemplo, no estudo da determinação da data de validade de produtos industrializados. A data de validade deve ser determinada de tal modo que o custo da reposição dos produtos inutilizados em dado período de tempo, seja por defeito de fabricação, seja por deterioração, possa ser absorvido pela indústria. Assim, interessa determinar o menor tempo no qual o fabricante possa suportar o custo dos produtos danificados.

Além da indústria, o tema em estudo pode ser aplicado também em outras áreas. Em finanças, por exemplo, pode ser de interesse determinar o tempo que um dado percentual de clientes de uma carteira se torna inadimplente. Nas Forças Armadas, é importante conhecer o tempo em que um percentual da munição ou da ração operacional do combatente falha, dadas algumas condições especiais de combate. Ou ainda, quantos quilômetros um pneu de carro de combate ou uma esteira de viatura

sobre lagartas pode rodar em terreno distinto para o qual foi projetado.

Na maior parte dessas aplicações, há interesse em distinguir o tempo associado a pequenas probabilidades de ocorrência dos eventos. Por isto, nesta dissertação será dada especial atenção aos quantis mais baixos da função quantílica.

1.2 Revisão da Literatura

O presente trabalho tem por finalidade estimar a função quantílica na presença de censura intervalar, buscando comparar o vício e a variabilidade das estimativas para diferentes tamanhos de amostra, padrões de intervalos de censura, métodos de determinação do parâmetro de suavização do estimador não paramétrico de máxima verossimilhança da função quantílica e ordem do quantil.

Para isto, os seguintes passos foram adotados:

i) estimação da função de distribuição marginal de T através do estimador não paramétrico de máxima verossimilhança (ENPMV), utilizando o algoritmo descrito em Groeneboom e Wellner (1992) e já implementado computacionalmente por Gomes (2006), bem como no pacote *Icens* da linguagem de programação *R*.

ii) suavização da estimativa não paramétrica de máxima verossimilhança da função de distribuição de T, utilizando núcleo estimador, e obtenção da função quantílica da variável aleatória T;

iii) realização de estudos de simulação, bem como aplicação a um conjunto de dados reais apresentado em Finkelstein e Wolfe (1985).

No Capítulo 2, são introduzidos conceitos básicos de análise de sobrevivência, no qual são também destacadas as funções básicas mais importantes empregadas. A teoria da *regressão isotônica* é exposta e algumas definições são apresentadas para dar suporte ao cálculo do *Estimador Não Paramétrico de Máxima Verossimilhança* (ENPMV). Barlow *et al.* (1972) mostram resultados importantes sobre este tema. Groeneboom e Wellner (1992) empregam a teoria da regressão isotônica para obter o ENPMV $\hat{F}$ de uma função de distribuição F.

Ao final do Capítulo 2, é apresentado um algoritmo desenvolvido por Groeneboom e Wellner (1992), e implementado computacionalmente por Gomes (2006), para calcular o ENPMV do caso Geral de Censura Intervalar. O algoritmo também está

implementado no pacote *Icens* da linguagem *R*.

No Capítulo 3 é apresentado o método do *núcleo estimador*, introduzido na literatura por Rosenblatt (1956) e Parzen (1962), para estimar, não parametricamente, a função de densidade de uma variável aleatória contínua, para dados completos. A determinação da *janela ótima* ou *parâmetro de suavização*, h, é descrita, bem como suas limitações.

Rudemo (1982) e Bowman (1984) propuseram o método de validação cruzada para determinar o valor ótimo de h para amostras não censuradas e Silverman (1986) faz uma excelente apresentação sobre este tema. Posteriormente, Chiu (1991) propõe uma modificação neste método para alcançar melhores resultados com a estabilização da variância.

Pan (2000) apresenta uma proposta para determinação da janela ótima na presença de censura intervalar. E, finalmente, Kulasekera e Padgett (2006) sugerem o emprego do critério de Bayes para estabelecer a janela ótima para dados censurados à direita.

Na Seção 3.5 são apresentados alguns estimadores da função quantílica.

Capítulo 2

Censura Intervalar

2.1 Introdução

Os estudos relativos à Análise de Sobrevivência reúnem um conjunto de técnicas e modelos estatísticos empregados para avaliar o comportamento de uma variável aleatória, geralmente o tempo até a ocorrência de um evento de interesse ("falha"). Uma característica comum nos dados é a presença de censura, o que ocorre quando os dados da variável resposta não são completamente observados.

Assim, são definidas duas variáveis aleatórias determinantes na análise de sobrevivência: o tempo de falha, T, e o tempo de censura, C. Quando se sabe apenas que $T > C$, tem-se censura à direita no instante C, significando que o evento de interesse ainda não ocorreu. A censura à esquerda se dá quando se sabe somente que $T < C$, indicando que a falha ocorreu anteriormente ao momento da observação, C.

Por outro lado, se o tempo de falha não é conhecido exatamente, sabendo-se apenas que ocorreu num dado intervalo de tempo $[U, V]$, tem-se respostas com censura intervalar. Esta estrutura de censura intervalar é conhecida como Caso 2 ou Caso Geral e é muito comum em estudos clínicos, nos quais os pacientes são acompanhados em consultas periódicas.

O Caso 1 de censura intervalar (ou dados de estado corrente) é mais simples que o Caso Geral e ocorre quando somente uma observação pode ser realizada sobre o estado da unidade amostral. Assim, no momento da verificação será observado se a falha já aconteceu (censura à esquerda) ou se ela ainda vai ocorrer (censura à direita).

Exemplificando, pode-se citar um experimento para estimar a distribuição de probabilidade do tempo até a deterioração de um produto alimentício, no qual a embalagem só pode ser aberta uma única vez para a verificação do estado do produto (se está deteriorado ou não).

A estrutura de censura intervalar é importante por que trata-se de uma generalização dos casos de censura à direita e à esquerda. De fato, tomando $U = C$ e $V = \infty$, obtém-se a censura à direita, e fazendo $U = 0$ e $V = C$, a censura à esquerda. Se $T = U = V$, observa-se o tempo exato de falha.

No estudo do comportamento da variável aleatória T, a *função de sobrevivência*, $S(t)$, tem um papel importante. Ela define a probabilidade da não ocorrência de falha até um dado tempo t, ou seja

$$S(t) = P(T \geq t).$$

A função de distribuição F de T é, portanto,

$$F(t) = 1 - S(t).$$

Outra função importante nos estudos de sobrevivência é a *taxa de falha instantânea*, $\lambda(t)$. É definida como o limite da probabilidade condicional de ocorrer uma falha no intervalo $[t, t + \Delta t]$, dado que não ocorreu antes de t, dividida pelo comprimento do intervalo, Δt, quando este tende a zero. É expressa por

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{\frac{S(t) - S(t+\Delta t)}{S(t)}}{\Delta t} = \lim_{\Delta t \to 0} \frac{S(t) - S(t + \Delta t)}{\Delta t S(t)} = \frac{f(t)}{S(t)}.$$

É fácil ver que $\lambda(t)$ é sempre positiva. Caso seja uma função crescente, o aumento da taxa de falha em função do tempo indica degradação cada vez maior da unidade amostral com o passar do tempo. Isto se dá no sentido de que a probabilidade condicional de ocorrer uma falha no intervalo $[t_1, t_2]$, dado que não ocorreu antes de t_1, será menor que a probabilidade condicional de ocorrer uma falha no intervalo $[t_2, t_3]$, dado que não ocorreu antes de t_2, $t_1 < t_2 < t_3$.

Por outro lado, a deterioração da unidade amostral é mais vagarosa ao longo do tempo se a taxa de falha é decrescente. Isto é, a probabilidade condicional de ocorrer uma falha no intervalo $[t_1, t_2]$, dado que não ocorreu antes de t_1, será maior que a

probabilidade condicional de ocorrer uma falha no intervalo $[t_2, t_3]$, dado que não ocorreu antes de t_2, $t_1 < t_2 < t_3$.

Já uma taxa de falha constante, denota a chamada falta de memória de $\lambda(t)$. Ou seja, a probabilidade de ocorrer falha após o tempo $t + s$, dado que não houve falha antes de t, é igual à probabilidade de ocorrer falha após o tempo s.

A *função taxa de falha acumulada*, embora não tenha uma interpretação fácil, é bastante útil na análise de sobrevivência. É definida por

$$\Lambda(t) = \int_0^t \lambda(u)du.$$

Outras duas quantidades importantes no estudo de análise de sobrevivência são o *tempo médio de vida*, t_m, e a *vida média residual*, $vmr(t)$. O tempo médio de vida, t_m, é obtido tomando-se o valor esperado do tempo de falha, T. Lembrando que T é uma variável não negativa, tem-se que

$$t_m = E(T) = \int_0^\infty (1 - F(t))\, dt = \int_0^\infty S(t)dt.$$

Já a vida média residual, $vmr(t)$, mede o tempo médio restante de vida após um dado tempo t (até ocorrer a falha). Ela está condicionada ao tempo t e é obtida por

$$vmr(t) = E(T - t|T > t) = \frac{\int_t^\infty (u - t)f(u)du}{S(t)} = \frac{\int_t^\infty S(u)du}{S(t)}.$$

Por fim, apresenta-se a *função quantílica*, que também é muito empregada em estudos de sobrevivência, e é definida por

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

A função quantílica, $Q(p)$, representa o menor tempo t em que pelo menos $100p\%$ das falhas ocorrem.

As funções relatadas estão relacionadas entre si e algumas dessas relações merecem destaque:

- $\diamond \quad -\frac{d}{dt}\ln S(t) = -\frac{d}{dt}\ln(1 - F(t)) = \frac{f(t)}{1-F(t)} = \frac{f(t)}{S(t)} = \lambda(t)$;
- $\diamond \quad \Lambda(t) = \int_0^t \lambda(u)du = -\ln S(t)$;
- $\diamond \quad S(t) = \exp\left[-\Lambda(t)\right] = \exp\left[-\int_0^t \lambda(u)du\right]$.

2.2 Função de Verossimilhança para Dados de Estado Corrente

Colosimo e Giolo (2006) ressaltam que todos os dados provenientes de um estudo de sobrevivência devem ser considerados na análise estatística, mesmo os censurados. Isto porque: (i) mesmo incompletas, as observações censuradas fornecem alguma informação sobre o tempo de falha; e (ii) a omissão de dados censurados no cálculo de estatísticas de interesse pode acarretar conclusões viciadas.

Dessa forma, no caso 1 de censura intervalar, por exemplo, deve-se considerar os pares de observações (c_i,δ_i). A variável C representa o tempo no qual se verifica o estado do indivíduo e a variável $\Delta = \mathrm{I}_{\{T \leq C\}}$ indica se a falha já ocorreu ou não. Ou seja, $\delta_i = 1$, se o tempo de falha do indivíduo i é menor ou igual ao tempo de censura, e $\delta_i = 0$, caso contrário.

Para obter a parte da função de verossimilhança que envolve apenas a função de distribuição F, de T, utiliza-se a distribuição conjunta de C e Δ. Assim, dado (c_i, δ_i), $i = 1, \ldots, n$, considerando G a função de distribuição e g a função densidade de probabilidade de C e assumindo T e C independentes, descreve-se a distribuição conjunta de C e Δ a seguir.

Para $\delta_i = 0$, tem-se que

$$\begin{aligned} P(C \leq c, \delta_i = 0) &= P(C \leq c, T > C) \\ &= \int_0^c \int_y^\infty f(t)g(y)dtdy \\ &= \int_0^c g(y) \int_y^\infty f(t)dtdy \\ &= \int_0^c g(y)[1 - F(y)]dy. \end{aligned}$$

E para $\delta_i = 1$, tem-se que

$$\begin{aligned} P(C \leq c, \delta_i = 1) &= P(C \leq c, T \leq C) \\ &= \int_0^c \int_0^y f(t)g(y)dtdy \\ &= \int_0^c g(y) \int_0^y f(t)dtdy \\ &= \int_0^c g(y)F(y)dy. \end{aligned}$$

Diferenciando as expressões acima com relação a c, tem-se que a densidade conjunta de C e Δ no ponto $(c,0)$ é igual a

$$\frac{d}{dc}\int_0^c g(y)[1-F(y)]dy = g(c)[1-F(c)],$$

e no ponto $(c,1)$ é igual a

$$\frac{d}{dc}\int_0^c g(y)F(y)dy = g(c)F(c).$$

Portanto, a função de verossimilhança é dada por

$$\begin{aligned} L(F) &= \prod_{i=1}^{n} \{g(c_i)F(c_i)\}^{\delta_i} \{g(c_i)[1-F(c_i)]\}^{1-\delta_i} \\ &= \prod_{i=1}^{n} g(c_i) \left[F(c_i)\right]^{\delta_i} \left[1-F(c_i)\right]^{1-\delta_i}. \end{aligned}$$

Assim,

$$L(F) \propto \prod_{i=1}^{n} \left[F(c_i)\right]^{\delta_i} \left[1-F(c_i)\right]^{1-\delta_i} \quad \text{e}$$

a função de log-verossimilhança é dada por

$$\mathcal{L}(F) = \sum_{i=1}^{n} \{\delta_i \ln F(c_i) + (1-\delta_i)\ln(1-F(c_i))\} + k, \tag{2.1}$$

onde k não depende de F.

Ao assumir-se um modelo paramétrico F_θ para a distribuição de T, substitui-se a expressão correspondente a F em (2.1) para encontrar o estimador de θ que maximiza $\mathcal{L}(\theta)$.

Quando não se tem idéia da forma de F, adota-se a abordagem não-paramétrica para solução do problema. Para calcular o Estimador Não Paramétrico de Máxima Verossimilhança (ENPMV) de F, é preciso encontrar $\hat{F}$ tal que $0 \leq \hat{F}(c_1) \leq \hat{F}(c_2) \leq \ldots \leq \hat{F}(c_n) \leq 1$ que maximiza $\mathcal{L}(F)$, assumindo, sem perda de generalidade, que $0 < c_1 < c_2 < \ldots < c_n$. Neste procedimento, emprega-se a teoria da regressão isotônica, que é discutida a seguir.

2.3 Regressão Isotônica

Seja $X = \{x_1, x_2, \ldots, x_k\}$, onde $x_1 < x_2 < \ldots < x_k$. Para $i = 1, 2, \ldots, k$, seja $y_j(x_i)$, $j = 1, 2, \ldots, n_i$, um conjunto de medidas de alguma quantidade, isto é, para

$x_i \in X$, $y_1(x_i), \ldots, y_{n_i}(x_i)$ são observações dependentes da distribuição de X. Seja $\mu(x_i)$ a esperança condicional de Y dado que $X = x_i$, isto é $\mu(x_i) = E(Y|X = x_i)$. Se $\mu(x_i)$ é uma função linear em X, pode ser de interesse estimá-la mediante o emprego de regressão linear, cuja solução é obtida pelo ajuste dos dados, minimizando

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}[y_j(x_i) - f(x_i)]^2 \tag{2.2}$$

na classe das funções lineares. Defina

$$\bar{y}(x_i) = \frac{1}{n_i}\sum_{j=1}^{n_i} y_j(x_i).$$

Sendo

$$\sum_{j=1}^{n_i}[y_j(x_i) - f(x_i)]^2 = \sum_{j=1}^{n_i}[y_j(x_i) - \bar{y}(x_i)]^2 + n_i[\bar{y}(x_i) - f(x_i)]^2,$$

o problema da minimização de (2.2) equivale a minimizar

$$\sum_{i=1}^{k} n_i[\bar{y}(x_i) - f(x_i)]^2. \tag{2.3}$$

Se não há restrição sobre μ, a solução obtida com a minimização de (2.3) é claramente a função f tal que $f(x_i) = \bar{y}(x_i)$, $i = 1, 2, \ldots, k$. Por outro lado, se $\mu(x_i)$ é não-decrescente com respeito a X, a estimativa de mínimos quadrados de μ vem da minimização da soma ponderada dos quadrados (2.3) na classe das funções não-decrescentes f. Esta solução pode ser chamada de regressão isotônica.

Definição 2.3.1. *Seja $X = \{x_1, \ldots, x_n\}$ com $x_1 \leq x_2 \leq \ldots \leq x_n$. Uma função $f(x) \to \Re$ é isotônica se para $x_i,\ x_j \in X,\ \forall\ i \neq j,\ x_i \leq x_j$ implica $f(x_i) \leq f(x_j)$.*

Definição 2.3.2. *Seja $X = \{x_1, \ldots, x_n\}$ com $x_1 \leq x_2 \leq \ldots \leq x_n$. Sejam g uma função em X, w uma função positiva em X e f uma função não decrescente com suporte em X. Uma função isotônica g^* em X é uma regressão isotônica de g com pesos w se g^* minimiza, na classe das funções isotônicas de X, a soma*

$$\sum_{x\in X}[g(x) - f(x)]^2 w(x),$$

isto é

$$g^* = \arg\min_{f\in\mathcal{F}}\{\sum_{x\in X}[g(x) - f(x)]^2 w(x)\},$$

onde $\mathcal{F}$ é a classe das funções isotônicas f definidas em X.

Barlow *et al.* (1972) mostram que a interpretação gráfica da regressão isotônica é obtida pela construção do *diagrama de somas acumuladas* (DSA), que é formado pelos pontos $(0,0)$ e $(\sum_{j=1}^{i} w(x_j), \sum_{j=1}^{i} w(x_j)g(x_j))$, $i = 1, \ldots, n$. A função g^* é dada pela derivada à esquerda da função minorante convexa máxima do DSA no ponto $\sum_{j=1}^{i} w(x_j)$.

Definição 2.3.3. *A função $f : (a,b) \to \Re$ é convexa se a primeira derivada de f é crescente em (a,b).*

Definição 2.3.4. *A função minorante convexa máxima do DSA formado pelos pontos $(0,0)$ e $(\sum_{j=1}^{i} w(x_j), \sum_{j=1}^{i} w(x_j)g(x_j))$, $i = 1, \ldots, n$ é a função*

$$H^* : [0, \sum_{j=1}^{n} w(x_j)] \to \Re$$

tal que

$$H^*(t) = \sup\{H(t) : H(\sum_{j \leq k} w(x_j)) \leq \sum_{j \leq k} w(x_j)g(x_j),\ 0 \leq k \leq n,\ H(0) = 0,\ H \text{ convexa}\}.$$

A Figura 2.1 ilustra um diagrama de somas acumuladas e sua função minorante convexa máxima.

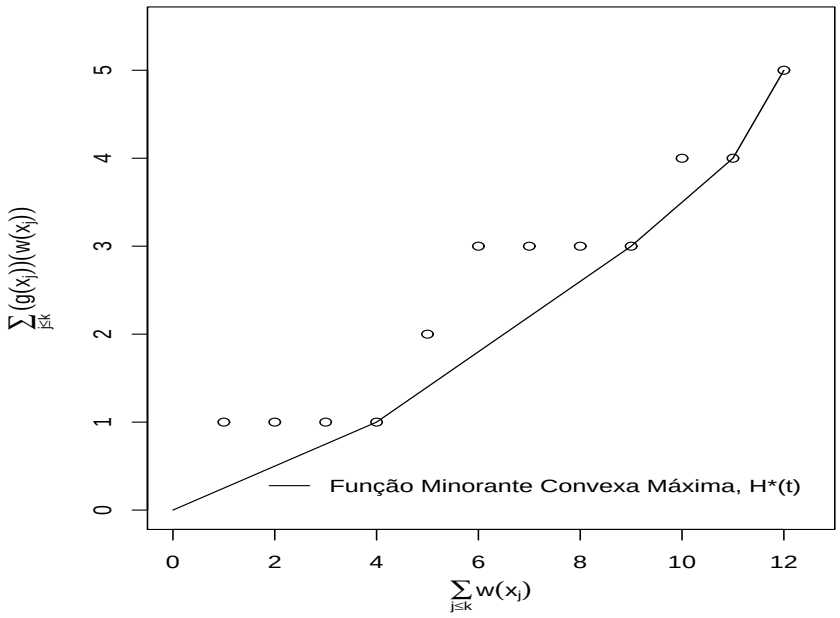


Figura 2.1: *Exemplo de uma função minorante convexa máxima.*

Os autores mostram que, alternativamente, obtém-se a regressão isotônica g^* empregando a seguinte fórmula:

$$g^*(x_i) = \max_{r \leq i} \min_{s \geq i} \frac{\sum_{m=r}^{s} g(x_m) w(x_m)}{\sum_{m=r}^{s} w(x_m)}. \tag{2.4}$$

A Tabela a seguir mostra os valores da regressão isotônica g^*, referentes aos dados apresentados na Figura 2.1.

Tabela 2.1: *Exemplo de regressão isotônica.*

$\sum_{j \leq k} w(x_j)$	$\sum_{j \leq k} w(x_j) g(x_j)$	$H^*\left(\sum_{j \leq k} w(x_j)\right)$	g^*
1	1	0.25	0.25
2	1	0.50	0.25
3	1	0.75	0.25
4	1	1.00	0.25
5	2	1.40	0.40
6	3	1.80	0.40
7	3	2.20	0.40
8	3	2.60	0.40
9	3	3.00	0.40
10	4	3.50	0.50
11	4	4.00	0.50
12	5	5.00	1.00

Ainda em Barlow *et al.* (1972), o teorema 1.10 mostra que a regressão isotônica g^* de g maximiza

$$\sum_{i=1}^{n} \{\Phi(f(x_i)) + [g(x_i) - f(x_i)]\phi(f(x_i))\} w(x_i) \tag{2.5}$$

na classe das funções isotônicas f, onde Φ é estritamente convexa e $\phi(y) = d\Phi(y)/dy$.

2.4 ENPMV - Estimador Não-Paramétrico de Máxima Verossimilhança

Como visto anteriormente, para calcular o estimador não-paramétrico de máxima verossimilhança (ENPMV) de F, no caso 1 de censura intervalar (quando apenas uma observação pode ser realizada sobre o estado da unidade amostral), é preciso encontrar $\hat{F}$ que maximize (2.1), tal que $0 \leq \hat{F}(c_1) \leq \hat{F}(c_2) \leq \ldots \leq \hat{F}(c_n) \leq 1$. Esta maximização será empreendida empregando-se a teoria de regressão isotônica. A função de log-verossimilhança dada em (2.1) pode ser escrita na forma da expressão (2.5), tomando

$$f = F, \quad x_i = c_i, \quad w(c_i) = 1, \quad g(c_i) = \delta_i$$

e fazendo

$$\Phi(F(c_i)) = F(c_i)\ln(F(c_i)) + (1 - F(c_i))\ln(1 - F(c_i)), \quad i = 1, \ldots, n.$$

Assim,

$$\begin{aligned} & \sum_{i=1}^{n} \{\Phi(F(c_i)) + [g(c_i) - F(c_i)]\phi(F(c_i))\} w(c_i) \\ = & \sum_{i=1}^{n} \{F(c_i)\ln F(c_i) + [1 - F(c_i)]\ln(1 - F(c_i)) \\ & + [\delta_i - F(c_i)][\ln F(c_i) - \ln(1 - F(c_i))]\} \\ = & \sum_{i=1}^{n} \{\delta_i \ln F(c_i) + (1 - \delta_i)\ln(1 - F(c_i))\} \\ = & \mathcal{L}(F). \end{aligned}$$

Portanto, o ENPMV $\hat{F}$ é tal que $\hat{F}(c_i)$, $i = 1, \ldots, n$, são dados pela regressão isotônica da função $g(c_i) = \delta_i$, com pesos $w(c_i) = 1$, sendo $\hat{F}(c_i)$ dado pela derivada à esquerda no ponto i do DSA formado pelos pontos

$$\left(\sum_{j=1}^{i} w(c_j), \sum_{j=1}^{i} w(c_j) g(c_j) \right) = \left(i, \sum_{j=1}^{i} \delta_j \right), \quad i = 1, \ldots, n. \tag{2.6}$$

Empregando (2.4) e considerando (2.6), pode-se calcular o ENPMV por

$$\hat{F}(c_i) = \max_{j \leq i} \min_{k \geq i} \frac{\sum_{m=j}^{k} \delta_m}{k - j + 1}.$$

2.5 Função de Verossimilhança do Caso Geral de Censura Intervalar

No caso geral de censura intervalar, as variáveis U, V, $\Delta = \mathrm{I}_{\{T \leq U\}}$ e $\Gamma = \mathrm{I}_{\{U < T \leq V\}}$ são observadas, onde $U \leq V$ com probabilidade 1. De maneira similar ao caso de dados de estado corrente, a função de verossimilhança é obtida a partir da distribuição conjunta de U, V, Δ e Γ. Sendo F a função de distribuição de T, H a função de distribuição conjunta de (U, V) e dados $(u_i, v_i, \delta_i, \gamma_i)$, $i = 1, \ldots, n$, descreve-se a distribuição conjunta de U, V, Δ e Γ a seguir.

Para $\delta_i = 1$ e $\gamma_i = 0$, tem-se que

$$\begin{aligned} P(U \leq u, V \leq v, \delta_i = 1, \gamma_i = 0) &= P(U \leq u, V \leq v, T \leq U) \\ &= \int_0^u \int_x^v \int_0^x f(t)h(x,y)dtdydx \\ &= \int_0^u \int_x^v h(x,y) \int_0^x f(t)dtdydx \\ &= \int_0^u \int_x^v h(x,y)F(x)dydx \\ &= \int_0^u F(x) \int_x^v h(x,y)dydx. \end{aligned}$$

Diferenciando com relação a u e v, obtém-se

$$\frac{\partial}{\partial u} \int_0^u F(x) \int_x^v h(x,y)dydx = F(u) \int_u^v h(u,y)dy$$

e

$$\begin{aligned} \frac{\partial^2}{\partial u \partial v} \int_0^u F(x) \int_x^v h(x,y)dydx &= \frac{d}{dv}F(u) \int_u^v h(u,y)dy \\ &= F(u)h(u,v). \end{aligned}$$

Para $\delta_i = 0$ e $\gamma_i = 1$, tem-se que

$$\begin{aligned} P(U \leq u, V \leq v, \delta_i = 0, \gamma_i = 1) &= P(U \leq u, V \leq v, U \leq T \leq V) \\ &= \int_0^u \int_x^v \int_x^y f(t)h(x,y)dtdydx \\ &= \int_0^u \int_x^v h(x,y) \int_x^y f(t)dtdydx \\ &= \int_0^u \int_x^v h(x,y)[F(y) - F(x)]dydx. \end{aligned}$$

Diferenciando com relação a u e v, obtém-se

$$\frac{\partial}{\partial u}\int_0^u\int_x^v h(x,y)[F(y)-F(x)]dydx = \int_u^v h(u,y)[F(y)-F(u)]dy$$

e

$$\begin{aligned}\frac{\partial^2}{\partial u\partial v}\int_0^u\int_x^v h(x,y)[F(y)-F(x)]dydx &= \frac{d}{dv}\int_u^v h(u,y)[F(y)-F(u)]dy\\ &= h(u,v)[F(v)-F(u)].\end{aligned}$$

Para $\delta_i = 0$ e $\gamma_i = 0$, tem-se que

$$\begin{aligned}P(U\leq u, V\leq v, \delta_i=0, \gamma_i=0) &= P(U\leq u, V\leq v, T>V)\\ &= \int_0^u\int_x^v\int_y^\infty f(t)h(x,y)dtdydx\\ &= \int_0^u\int_x^v h(x,y)\int_y^\infty f(t)dtdydx\\ &= \int_0^u\int_x^v h(x,y)[1-F(y)]dydx.\end{aligned}$$

Diferenciando com relação a u e v, obtém-se

$$\frac{\partial}{\partial u}\int_0^u\int_x^v h(x,y)[1-F(y)]dydx = \int_u^v h(u,y)[1-F(y)]dy$$

e

$$\begin{aligned}\frac{\partial^2}{\partial u\partial v}\int_0^u\int_x^v h(x,y)[1-F(y)]dydx &= \frac{d}{dv}\int_u^v h(u,y)[1-F(y)]dy\\ &= h(u,v)[1-F(v)].\end{aligned}$$

Para $\delta_i = 1$ e $\gamma_i = 1$, não há massa de probabilidade.

Portanto, a função de verossimilhança é dada por

$$\begin{aligned}L(F) &= \prod_{i=1}^n [h(u_i,v_i)F(u_i)]^{\delta_i}[h(u_i,v_i)(F(v_i)-F(u_i))]^{\gamma_i}[h(u_i,v_i)(1-F(v_i))]^{(1-\delta_i-\gamma_i)}\\ &= \prod_{i=1}^n h(u_i,v_i)[F(u_i)]^{\delta_i}[F(v_i)-F(u_i)]^{\gamma_i}[1-F(v_i)]^{(1-\delta_i-\gamma_i)}.\end{aligned}$$

Dessa forma,

$$L(F) \propto \prod_{i=1}^n [F(u_i)]^{\delta_i}[F(v_i)-F(u_i)]^{\gamma_i}[1-F(v_i)]^{(1-\delta_i-\gamma_i)},$$

e a função de log-verossimilhança é dada por

$$\mathcal{L}(F) = \sum_{i=1}^n \delta_i log([F(u_i)]) + \gamma_i log([F(v_i)-F(u_i)]) + (1-\delta_i-\gamma_i)log([1-F(v_i)]) + k,$$

onde k não depende de F.

2.6 ENPMV do Caso Geral de Censura Intervalar

Groeneboom e Wellner (1992) introduziram o *algoritmo do minorante convexo máximo*, mais conhecido como algoritmo ICM (*Iterative Convex Minorant*), para obter o ENPMV.

Empregando resultados da regressão isotônica, os autores mostram que o ENPMV $\hat{F}$ pode ser obtido pela derivada à esquerda da função minorante convexa máxima de um Diagrama de Somas Acumuladas.

Assim, Groeneboom e Wellner (1992) demonstram o seguinte teorema:

Teorema 2.6.1. *Seja M_1 correspondente a uma observação U_i tal que $I_{\{T_i \leq U_i\}} = 1$ e seja a maior estatística de ordem M_m correspondente a uma observação V_i tal que $I_{\{T_i \geq V_i\}} = 1$. Então, $\hat{F}$ é o ENPMV de F se e somente se $\hat{F}$ é a derivada à esquerda da função minorante convexa máxima do DSA, que é formado pelos pontos $P_j = (G_{\hat{F}}(M_j), H_{\hat{F}}(M_j))$, $j = 1, \ldots, m$, e $P_0 = (0, 0)$.*

As funções $G_F(t)$ e $H_F(t)$ são definidas como se segue:

$$G_F(t) = \frac{1}{n}\left\{\sum_{U_i \leq t} \frac{\delta_i}{[F(U_i)]^2} + \sum_{U_i \leq t} \frac{\gamma_i}{[F(V_i) - F(U_i)]^2} + \sum_{V_i \leq t} \frac{\gamma_i}{[F(V_i) - F(U_i)]^2} + \sum_{V_i \leq t} \frac{1 - \delta_i - \gamma_i}{[1 - F(V_i)]^2}\right\}$$

e

$$H_F(t) = W_F(t) + \sum_{M_j \leq t} F(M_j)[G_F(M_j) - G_F(M_{j-1})],$$

sendo,

$$W_F(t) = \frac{1}{n}\left\{\sum_{U_i \leq t} \frac{\delta_i}{F(U_i)} - \sum_{U_i \leq t} \frac{\gamma_i}{F(V_i) - F(U_i)} + \sum_{V_i \leq t} \frac{\gamma_i}{F(V_i) - F(U_i)} - \sum_{V_i \leq t} \frac{1 - \delta_i - \gamma_i}{1 - F(V_i)}\right\}$$

e os M_j, $j = 1, \ldots, m$, os valores ordenados do conjunto

$$\begin{aligned} J = {} & \{U_i : \delta_i = 1 \text{ ou } \gamma_i = 1, \quad i = 1, \ldots, n\} \cup \\ & \{V_i : \gamma_i = 1 \text{ ou } \delta_i = \gamma_i = 0, \quad i = 1, \ldots, n\}, \end{aligned} \tag{2.7}$$

$M_0 = 0$, $G_F(0) = 0$ e $m = n + \sum_{i=1}^{n} \gamma_i$.

Note que os pontos do DSA dependem de $\hat{F}$. Portanto, tem-se que adotar um algoritmo iterativo para obtenção de $\hat{F}$.

O *algoritmo iterativo do minorante convexo* empregado para calcular o ENPMV de $\hat{F}$, descrito em Groeneboom e Wellner (1992), consiste nos seguintes passos:

i) Tome $F^{(0)}(M_j) = \frac{j}{m}$, $j = 1, \ldots, m$.

ii) Construa o DSA com os pontos $P_0 = (0, 0)$ e $P_j = (G_{\hat{F}^{(k)}}(M_j), H_{\hat{F}^{(k)}}(M_j))$, $j = 1, \ldots, m$ e obtenha $F^{(k+1)}(M_j)$ como sendo a derivada à esquerda no ponto $G_{\hat{F}^{(k)}}(M_j)$ da função minorante convexa máxima do DSA.

iii) Critério de parada: $|| F^{(k+1)} - F^{(k)} || < \varepsilon$, para alguma norma $|| \cdot ||$.

Este algoritmo será empregado neste trabalho e foi implementado computacionalmente por Gomes (2006), estando também disponível no pacote *Icens* da linguagem *R* (disponível em <http://cran-r.c3sl.ufpr.br>).

Capítulo 3

Núcleo Estimador

3.1 Introdução

Introduzido na literatura por Rosenblatt (1956) e Parzen (1962), o método do núcleo estimador procura estimar, não parametricamente, a função densidade de uma variável aleatória X. Dada uma amostra aleatória $X_1, \ldots, X_n$, o estimador de f 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 janela ou parâmetro de suavização e $K(\cdot)$ é uma função densidade de probabilidade qualquer, chamada núcleo. Normalmente são empregadas formas usuais não negativas e simétricas para a função K, que está relacionada à forma de $\hat{f}$. A janela h relaciona-se com o grau de suavização e sua determinação é, geralmente, mais importante do que a escolha de K para obter um bom desempenho do método (Silverman, 1986).

3.2 Estimação Para Dados Não Censurados

Antes de considerar os dados censurados, objetivo maior deste estudo, é interessante fazer algumas colocações sobre o núcleo estimador para dados completos. Duas importantes propriedades do núcleo estimador merecem destaque:

(i) Se a função $K(\cdot)$ é não negativa e satisfaz

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

então $\hat{f}$ é uma função densidade;

(ii) Todas as propriedades locais de continuidade e diferenciabilidade de $K(\cdot)$ são também herdadas por $\hat{f}$ (Silverman, 1986).

Algumas propostas têm sido estudadas para medir a discrepância entre o estimador $\hat{f}$ e f, a verdadeira e desconhecida função de densidade. A medida natural de discrepância entre $\hat{f}$ e f, considerando a estimação pontual, é o erro quadrático médio

$$EQM_x(\hat{f}) = E\{(\hat{f}(x) - f(x)\}^2. \tag{3.1}$$

Utilizando propriedades elementares de média e variância, é possível reescrever (3.1) como

$$EQM_x(\hat{f}) = \{E\hat{f}(x) - f(x)\}^2 \quad + \quad Var\hat{f}(x). \tag{3.2}$$

Isto é, o erro quadrático médio pode ser expresso pela soma da variância e do quadrado do vício.

Uma medida global da acurácia de $\hat{f}$, como um estimador de f, pode ser obtida pelo erro quadrático médio integrado, definido por

$$EQMI(\hat{f}) = E\int\{\hat{f}(x) - f(x)\}^2 dx. \tag{3.3}$$

Como o integrando é não negativo, é possível inverter a ordem da integração e da esperança em (3.3), resultando em

$$\begin{aligned} EQMI(\hat{f}) &= \int E\{\hat{f}(x) - f(x)\}^2 dx \\ &= \int EQM_x(\hat{f})dx \\ &= \int\{E\hat{f}(x) - f(x)\}^2 dx \quad + \quad \int Var\hat{f}(x)dx, \end{aligned} \tag{3.4}$$

de modo que o EQMI pode ser escrito como a soma da integral do quadrado do vício e da integral da variância.

3.3 Estimação Para Dados Censurados

Para estimar f no caso de dados censurados, o estimador não-paramétrico comumente utilizado é dado por

$$\hat{f}(x) = \frac{1}{h}\int_{-\infty}^{\infty} K\left(\frac{x-t}{h}\right) d\hat{F}(t), \tag{3.5}$$

onde $\hat{F}$ é o estimador da função de distribuição da variável que descreve o tempo de vida, $K(\cdot)$ é a função núcleo e h é o parâmetro de suavização.

Para dados com censura intervalar, o estimador da densidade em (3.5) pode ser reescrito como

$$\hat{f}(x) = \frac{1}{h}\sum_{j=1}^{n} s_j K\left(\frac{x - M_j}{h}\right), \tag{3.6}$$

onde $s_j = \hat{S}(M_j) - \hat{S}(M_{j+1}) = \hat{F}(M_{j+1}) - \hat{F}(M_j)$, os M_j são os valores ordenados do conjunto J em (2.7) e $\hat{F}$ é calculado da maneira especificada na Seção 2.6.

3.4 Determinação da Janela Ótima

A determinação da janela ótima é descrita por Silverman (1986) a partir de uma função de dois argumentos $w(y, x)$, satisfazendo às seguintes condições

$$\int_{-\infty}^{\infty} w(y, x)dx = 1 \tag{3.7}$$

e

$$w(y, x) \geq 0. \tag{3.8}$$

É possível definir $\hat{f}$ numa classe geral de estimadores de densidade

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} w(Y_i, x), \tag{3.9}$$

desde que satisfeitas as condições (3.7) e (3.8).

Observe que

$$E\left[\hat{f}(x)\right] = \frac{1}{n}\sum_{i=1}^{n} Ew(Y_i, x) = \int_{-\infty}^{\infty} w(y, x)f(y)dy \tag{3.10}$$

e que, para Y_i independentes,

$$\begin{aligned} Var\left[\hat{f}(x)\right] &= \frac{1}{n}Var\left[w(Y_i, x)\right] \\ &= \frac{1}{n}\left\{\int w(y, x)^2 f(y)dy - \left[\int w(y, x)f(y)dy\right]^2\right\}. \end{aligned} \tag{3.11}$$

O núcleo estimador pode ser obtido fazendo

$$w(y, x) = \frac{1}{h}K\left(\frac{x - y}{h}\right). \tag{3.12}$$

Substituindo (3.12) em (3.10) e em (3.11), tem-se

$$E\left[\hat{f}(x)\right] = \int \frac{1}{h} K\left(\frac{x-y}{h}\right) f(y)dy \tag{3.13}$$

e

$$Var\left[\hat{f}(x)\right] = \frac{1}{n}\int \frac{1}{h^2} K\left(\frac{x-y}{h}\right)^2 f(y)dy - \left[\frac{1}{h}\int K\left(\frac{x-y}{h}\right) f(y)dy\right]^2. \tag{3.14}$$

Para a determinação da janela ótima, suponha que K é uma função simétrica satisfazendo

$$\int K(t)dt = 1, \quad \int tK(t)dt = 0 \quad \text{e} \quad \int t^2 K(t)dt = k_2 > 0$$

e que a densidade f tem derivadas contínuas de todas as ordens requeridas.

Considerando (3.13), pode-se reescrever o vício da estimação, $b(\hat{f}(x))$, como

$$\begin{aligned} b(\hat{f}(x)) &= E\left[\hat{f}(x)\right] - f(x) \\ &= \int \frac{1}{h} K\left(\frac{x-y}{h}\right) f(y)dy - f(x). \end{aligned} \tag{3.15}$$

Em geral, o vício e a variância não são determinados exatamente com facilidade. Silverman (1986) propõe uma aproximação para essas quantidades, empregando a expansão em série de Taylor.

Inicialmente, para encontrar uma aproximação para o vício, Silverman (1986) sugere uma mudança de variável em (3.15), fazendo $y = x - ht$. Considerando as suposições acima, tem-se, então,

$$\begin{aligned} b(\hat{f}(x)) &= \int K(t) f(x-ht)dt - f(x) \\ &= \int K(t)\left[f(x-ht) - f(x)\right] dt. \end{aligned}$$

A expansão em série de Taylor de $f(x-ht)$ em torno de x resulta

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

Assim, obtém-se a seguinte expressão para o vício

$$\begin{aligned} b(\hat{f}(x)) &= -hf'(x)\int tK(t)dt + \frac{1}{2}h^2 f''(x)\int t^2 K(t)dt + \ldots \\ &= \frac{1}{2}h^2 f''(x)k_2 \quad + \quad O(h). \end{aligned}$$

Dessa forma, a primeira parcela da soma em (3.4) é dada por

$$\int \{b(\hat{f}(x))\}^2 dx \approx \frac{1}{4} h^4 k_2^2 \int [f''(x)]^2 dx.$$

A obtenção da variância ocorre de modo semelhante, com suposição adicional de que h é pequeno e n é grande. Assim, a segunda parcela de (3.4) é

$$\int Var[\hat{f}(x)] dx \approx \frac{1}{nh} \int [K(t)]^2 dt.$$

Portanto, a janela ótima é obtida pela minimização de $EQMI(\hat{f})$ em função de h, resultando

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

3.4.1 Método da Validação Cruzada

A determinação da janela ótima, obtida acima, é um pouco desapontadora, no sentido de que ela depende da derivada de segunda ordem da desconhecida função de densidade. Rudemo (1982) e Bowman (1984) propuseram o método da validação cruzada para amostras não censuradas, objetivando estimar f.

A idéia consiste em minimizar o Erro Quadrático Integrado (EQI), que pode ser escrito como

$$\int (\hat{f} - f)^2 \;\; = \;\; \int \hat{f}^2 - 2 \int \hat{f} f + \int f^2, \tag{3.16}$$

onde $\hat{f}$ estima f.

Note que, para minimizar o EQI em função de h, basta minimizar as duas primeiras parcelas do segundo membro de (3.16)

$$R(\hat{f}) = \int \hat{f}^2 - 2 \int \hat{f} f. \tag{3.17}$$

O método da validação cruzada propõe construir uma estimativa para a quantidade $R(\hat{f})$ a partir dos dados observados. Para isto, considera estimar f por uma combinação das funções $\hat{f}_{-i}$, onde $\hat{f}_{-i}$ é a densidade estimada construída por todos os pontos exceto X_i, a qual também é conhecida como versão "menos um" de $\hat{f}$.

Dessa forma, Silverman (1986) define

$$M_0(h) = \int \hat{f}^2 - 2n^{-1} \sum_{i=1}^{n} \hat{f}_{-i}(X_i),$$

que depende somente dos dados.

A densidade $\hat{f}_{-i}$, segundo o autor, é dada por

$$\hat{f}_{-i}(x) = (n-1)^{-1}h^{-1}\sum_{j\neq i} K\{h^{-1}(x - X_j)\}. \tag{3.18}$$

O método da validação cruzada consiste em minimizar $M_0(h)$, em função de h. Silverman (1986) mostra que a justificativa para este procedimento encontra-se no fato de que $E[R(\hat{f})] = E[M_0(h)]$. Portanto, assumindo que o mínimo de $M_0(h)$ está próximo do mínimo de $E[M_0(h)]$, espera-se obter uma boa estimativa do parâmetro de suavização.

Por facilidade computacional, o autor sugere uma aproximação para $M_0(h)$, $M_1(h)$, obtida substituindo-se o fator $(n-1)^{-1}$ por n^{-1} em (3.18). O valor ótimo de h é o valor que minimiza $M_1(h)$.

Um resultado importante, que justifica o emprego deste método, foi dado por Stone (1984). Dada uma amostra aleatória $X_1, \ldots, X_n$ de uma densidade f, seja $EQI_{M_1(h)}$ o EQI da densidade estimada construída usando o parâmetro de suavização que minimiza a função $M_1(h)$. Seja EQI_{otimo} o menor EQI sobre todo h, mantendo os dados fixos. Sob condições muito fracas (Silverman, 1986), Stone (1984) mostra que

$$\frac{EQI_{M_1(h)}}{EQI_{otimo}} \to 1, \quad \text{quando} \quad n \to \infty. \tag{3.19}$$

3.4.2 Método da Validação Cruzada Modificado

O método de validação cruzada obtém estimativas de h com muita variabilidade. Visando obter a estabilização da variância, Chiu (1991) propôs uma expressão aproximada para a janela obtida pelo método de validação cruzada baseada na função característica.

No Método da Validação Cruzada Modificado, o h ótimo é o valor que minimiza

$$S_n(h) = \pi(nh)^{-1}\int K^2(x)dx + \int_0^{\Lambda}\left\{\left|\tilde{\phi}(\lambda)\right|^2 - n^{-1}\right\}\left\{W^2(h\lambda) - 2W(h\lambda)\right\}d\lambda,$$

onde

$$W(\lambda) = \int \exp(i\lambda x)K(x)dx,\ \Lambda = \min\{\lambda : \left|\tilde{\phi}(\lambda)\right|^2 \leq c/n\},$$

para alguma constante $c > 1$, e $\tilde{\phi}(\lambda)$ é a função característica amostral definida por

$$\tilde{\phi}(\lambda) = n^{-1} \sum_{j=1}^{n} \exp(i\lambda X_j), \ -\infty < \lambda < \infty, \ i = \sqrt{-1}.$$

3.4.3 Método para Dados Censurados à Direita

Sejam $X_1, \ldots, X_n$ variáveis aleatórias independentes e identicamente distribuídas (iid) denotando o tempo de falha, e $U_1, \ldots, U_n$ variáveis aleatórias iid, independentes dos $X_i's$, tais que são observados $Y_i = \min\{X_i, U_i\}$ e $\Delta_i = \mathrm{I}_{\{X_i \leq U_i\}}$. Defina as funções de distribuição dos $X_i's$ e dos $U_i's$ por F e H, respectivamente, e a função densidade dos $X_i's$ por f. Sejam $S = 1 - F$ e $H^* = 1 - H$. Baseado em (Y_i, Δ_i), $i = 1, \ldots, n$, considere o estimador da função de sobrevivência mais comumente empregado, $\hat{S}_n(t)$, proposto por Kaplan e Meier (1958). Sendo (Z_i, Λ_i), $i = 1, \ldots, n$, os $Y_i's$ ordenados com seus correspondentes $\Delta_i's$, o estimador de Kaplan-Meier é definido por

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

O estimador de f, empregando o método do núcleo estimador, pode ser dado por

$$\tilde{f}(x) = \frac{1}{h} \sum_{j=1}^{n} s_j K\left(\frac{x - Z_j}{h}\right),$$

onde s_j é o salto de $\hat{S}_n$ no ponto Z_j, definido por

$$\hat{s}_j = \begin{cases} 1 - \hat{S}_n(Z_2), & j = 1, \\ \hat{S}_n(Z_j) - \hat{S}_n(Z_{j+1}), & j = 2, \ldots, n-1, \\ \hat{S}_n(Z_n), & j = n. \end{cases}$$

Blum e Susarla (1980, *apud* Marron e Padgett, 1987) propuseram um estimador não-paramétrico alternativo para a função densidade, a partir de resultados obtidos por Rosemblatt (1976), mediante o emprego da função núcleo. Sua motivação se deu pelo fato de que uma razoável estimativa de $f(x)H^*(x)$ é dada por

$$(fH^*)_n(x) = \frac{1}{nh} \sum_{j=1}^{n} K\left(\frac{x - X_j}{h}\right) 1_{[\Delta_j = 1]}.$$

Logo, é possível estimar $f(x)$ por $(fH^*)_n(x)$ dividido por uma estimativa de $H^*(x)$.

Citando Blum e Susarla (1980), Marron e Padgett (1987) mostram que tomando o estimador de H^* obtido pela inversão dos papéis dos $X_i's$ e $U_i's$

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

obtém-se

$$\check{f}(x) = \frac{1}{nhH_n^*(x)} \sum_{j=1}^{n} K\left(\frac{x - X_j}{h}\right) 1_{[\Delta_j = 1]}.$$

Os autores mostram que tomando $s_j = \Lambda_j[n\hat{H}_n(Z_j)]^{-1}$, tem-se, então, os seguintes estimadores de f

$$\tilde{f}(x) = \sum_{j=1}^{n} \frac{\Delta_j}{n\hat{H}_n(X_j)h} K\left(\frac{x - X_j}{h}\right), \tag{3.20}$$

e

$$\check{f}(x) = \sum_{j=1}^{n} \frac{\Delta_j}{nH_n^*(x)h} K\left(\frac{x - X_j}{h}\right). \tag{3.21}$$

Desde que $\hat{H}_n$ e H_n^* são essencialmente iguais, a única diferença significante entre os estimadores é o argumento da estimativa de H^*.

Com base nesses resultados, Marron e Padgett (1987) utilizam o

$$\text{EQI}(\hat{f}) = \int_0^\infty [\hat{f}(x) - f(x)]^2 w(x) dx,$$

onde $w(x)$ é uma função não-negativa de "pesos", para avaliar a performance dos estimadores. Tal como em (3.16), a minimização do EQI implica em minimizar

$$\int_0^\infty [\hat{f}(x)^2 - 2\hat{f}(x)f(x)]w(x)dx.$$

Os autores buscam minimizar esta quantidade, considerando a versão "menos um" dos estimadores de f definidos em (3.20) e (3.21).

Assim, a janela ótima é o valor de h que minimiza a expressão

$$CV(h) = \int [\hat{f}(x)]^2 w(x)dx - 2n^{-1} \sum_i \hat{f}_{-i}(X_i) \frac{w(X_i)}{\hat{H}_n(X_i)}. \tag{3.22}$$

Os autores mostram ainda que, como em (3.19),

$$\frac{EQI_{CV(h)}}{EQI_{otimo}} \to 1, \quad \text{quando} \quad n \to \infty.$$

3.4.4 Método na Presença de Censura Intervalar

Um método alternativo foi proposto por Pan (2000) para escolha da janela ótima na presença de censura intervalar. Segundo o autor, o valor ótimo de h é dado por

$$h_{otimo} = \arg\max_{h} \sum_{\upsilon=1}^{V} L(\tilde{S}^{(-\upsilon)}(\cdot; h)|D^{(\upsilon)}),$$

onde

$$L(S|D) = \sum_{i=1}^{n} [\delta_i \ln(1 - S(c_i)) + (1 - \delta_i) \ln S(c_i)]$$

é a função de log-verossimilhança do caso 1 de censura intervalar e $\tilde{S}^{(-\upsilon)}(\cdot; h)$ é a função de sobrevivência estimada pelo método do núcleo, a partir das observações em $D \backslash D^{(\upsilon)}$, $\upsilon = 1, \ldots, V$, onde D representa o conjunto de dados subdividido aleatoriamente em V subconjuntos de mesmo tamanho. Isto é, $D = \{D^{(1)}, \ldots, D^{(\upsilon)}\}$.

Neste trabalho, $L(S|D)$ será a função de log-verossimilhança do caso geral de censura intervalar,

$$L(S|D) = \sum_{i=1}^{n} \delta_i \ln[F(u_i)] + \gamma_i \ln[F(v_i) - F(u_i)] + (1 - \delta_i - \gamma_i) \ln[1 - F(v_i)].$$

3.5 Estimação da Função Quantílica

Como visto anteriormente, a função quantílica $Q(p)$ representa o menor tempo t em que pelo menos $100p\%$ das falhas ocorrem e é definida por

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

Neste trabalho, serão empregados quatro estimadores da função quantílica. O primeiro deles será dado pela inversão do ENPMV $\hat{F}$ de F, obtendo

$$\hat{Q}_n(p) = \inf\{t : \hat{F}_n(t) \geq p\}, \quad 0 \leq p \leq 1.$$

O segundo estimador será obtido invertendo o ENPMV $\hat{F}$ de F, após suavizá-lo pelo método do núcleo estimador, resultando

$$\tilde{Q}_n^C(p) = \inf\{t : \tilde{F}_n(t) \geq p\}, \quad 0 \leq p \leq 1,$$

onde,

$$\begin{aligned}\tilde{F}_n(t) &= \int_{-\infty}^{\infty} \mathcal{K}\left(\frac{t-x}{h}\right) d\hat{F}_n(x) \\ &= \sum_{j=1}^{m} \mathcal{K}\left(\frac{t-M_j}{h}\right)\left(\hat{F}_n(M_j) - \hat{F}_n(M_{j-1})\right),\end{aligned}$$

sendo $\mathcal{K}$ uma função de distribuição.

A função núcleo empregada será a gaussiana. Para obtenção da janela ótima nesta suavização, será empregado o método de validação cruzada modificado (Chiu, 1991).

O terceiro estimador será semelhante ao segundo, alterando apenas a forma de obtenção da janela ótima para a suavização. Neste caso, será empregado o método de Pan (2000) para dados com censura intervalar e o estimador será denominado $\tilde{Q}_n^P(p)$.

Finalmente, o quarto estimador de $Q(p)$ será obtido a partir da suavização de $\hat{Q}_n(p)$ pelo método do núcleo estimador, empregando a função gaussiana. Para dados de censura intervalar, este estimador é definido por

$$\begin{aligned}\check{Q}_n(p) &= \int_0^1 \mathcal{K}\left(\frac{p-t}{h}\right) d\hat{Q}_n(t) \\ &= \sum_{j=1}^{m} \mathcal{K}\left(\frac{p-p_j}{h}\right)\left[\hat{Q}_n(p_j) - \hat{Q}_n(p_{j-1})\right], \qquad (3.23)\end{aligned}$$

onde $p_j, \quad j = 1, \ldots, m$, são os pontos de salto de $\hat{Q}_n(p)$.

O método de validação cruzada modificado (Chiu, 1991) será empregado para obtenção da janela ótima, nesta suavização.

A Figura 3.1 ilustra os estimadores da função quantílica $Q(p)$, sendo Q_{Real} a função quantílica real e Q_1, Q_2, Q_3 e Q_4, respectivamente, o primeiro, segundo, terceiro e quarto estimadores.

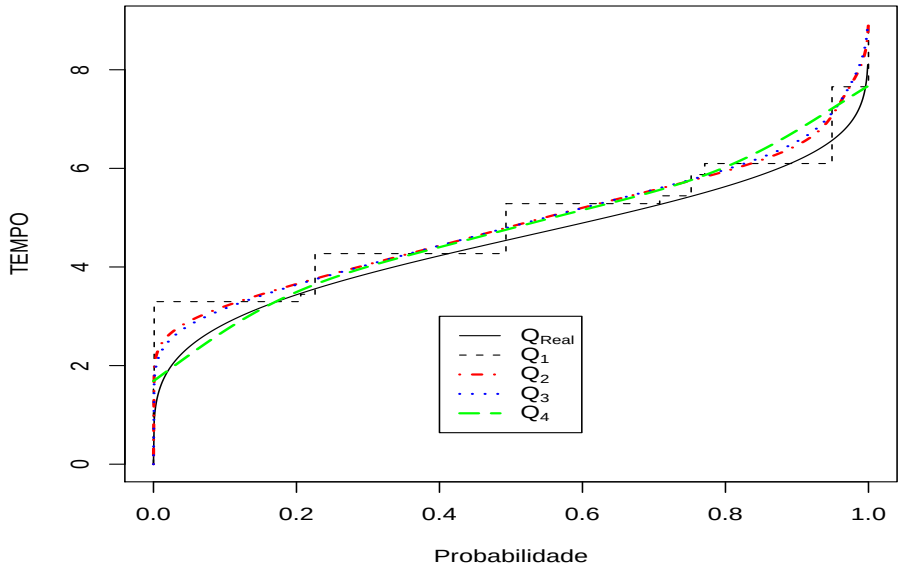


Figura 3.1: *Comparação entre os estimadores da função quantílica.*

Capítulo 4

Simulação e Aplicação

4.1 Simulação

Para os estudos de simulação realizados, foi empregada a distribuição Weibull, cujas funções densidade de probabilidade e de distribuição são dadas, respectivamente, por

$$f(t) = \frac{\gamma}{\alpha^{\gamma}} t^{\gamma-1} \exp\left\{-\left(\frac{t}{\alpha}\right)^{\gamma}\right\}, \quad t \geq 0,$$

e

$$F(t) = 1 - \exp\left\{-\left(\frac{t}{\alpha}\right)^{\gamma}\right\},$$

onde γ é o parâmetro de forma e α é o de escala.

Nas simulações, o tempo de falha T segue uma distribuição de Weibull com parâmetros $\gamma = 4$ e $\alpha = 5$.

Foram considerados, também, quatro diferentes padrões de intervalo de censura intervalar, mediante a definição de distintos parâmetros para as distribuições das variáveis U e V. Estas variáveis, como dito anteriormente, definem o intervalo de tempo no qual ocorreu a falha e, em consequência, o percentual de falhas que foram censuradas à esquerda em U, no intervalo $(U, V]$ e à direita em V. A Tabela (4.1) apresenta estas configurações.

Os estimadores da função quantílica presentes no estudo são os definidos na Seção 3.5, como se segue:

i) $Q_1 = \hat{Q}_n(p)$;

Tabela 4.1: *Configurações de intervalo de censura intervalar (C), definidas mediante a distribuição de U e $V-U$.*

	distribuição		percentual de falhas		
C	U	$V-U$	$(0,U]$	$(U,V]$	$(V,+\infty)$
1	Weibull $(\gamma=4,\alpha=3.8)$	Weibull $(\gamma=4,\alpha=2.5)$	25%	50%	25%
2	Weibull $(\gamma=4,\alpha=4.5)$	Weibull $(\gamma=4,\alpha=1.0)$	40%	20%	40%
3	Weibull $(\gamma=4,\alpha=3.8)$	Weibull $(\gamma=4,\alpha=1.2)$	25%	25%	50%
4	Weibull $(\gamma=4,\alpha=5.0)$	Weibull $(\gamma=4,\alpha=1.2)$	50%	25%	25%

ii) $Q_2 = \tilde{Q}_n^C(p)$, empregando o método de validação cruzada modificado (Chiu, 1991) para obtenção da janela ótima na suavização do ENPMV $\hat{F}$;

iii) $Q_3 = \tilde{Q}_n^P(p)$, empregando o método de Pan (2000) para dados com censura intervalar na obtenção da janela ótima na suavização do ENPMV $\hat{F}$;

iv) $Q_4 = \check{Q}_n(p)$, empregando o método de validação cruzada modificado (Chiu, 1991) para obtenção da janela ótima na suavização de $Q_1 = \hat{Q}_n(p)$.

Os quantis analisados foram: $q_1 = 0.01$, $q_2 = 0.02$, $q_3 = 0.05$, $q_4 = 0.10$, $q_5 = 0.20$ e $q_6 = 0.50$.

Foram realizadas simulações para quatro tamanhos de amostra: 50, 100, 200 e 500.

4.1.1 Vício dos Estimadores

Da análise da Tabela 4.2, percebe-se que o vício de Q_1 decresce à medida que aumentam o tamanho da amostra e o quantil, para todas as configurações de intervalo de censura, de modo geral. Os quantis são superestimados por Q_1 em todas as configurações e o tamanho do vício se mostra um pouco mais elevado para as configurações 2 e 4, nas quais as falhas t_i ocorrem mais intensamente no intervalo $(0,U)$.

Também é possível comparar a magnitude do vício na Tabela 4.3, a qual revela ainda que o vício relativo decresce à medida que aumentam o quantil e o tamanho da amostra.

Tabela 4.2: *Vício do estimador Q_1, segundo a configuração (C) e o tamanho da amostra (n).*

		Quantis					
C	n	0.01	0.02	0.05	0.10	0.20	0.50
1	50	1.225824	0.923900	0.467065	0.228333	0.077632	0.012304
	100	0.969158	0.667235	0.281642	0.115194	0.052626	-0.001577
	200	0.769405	0.478253	0.180007	0.067613	0.023153	0.005061
	500	0.501438	0.264450	0.080980	0.034651	0.024781	-0.002513
2	50	1.377303	1.075380	0.593736	0.287092	0.072880	-0.009171
	100	1.117362	0.815438	0.367382	0.138475	0.058791	-0.006852
	200	0.920251	0.620235	0.267194	0.092925	0.020508	0.008563
	500	0.644385	0.365651	0.105365	0.023093	0.025904	0.003048
3	50	1.164216	0.862293	0.418115	0.197530	0.049740	0.015872
	100	0.928595	0.627598	0.252809	0.077475	0.035054	0.000293
	200	0.738078	0.447123	0.150277	0.047705	0.019184	0.009925
	500	0.475367	0.240724	0.049749	0.026114	0.022396	0.003277
4	50	1.537493	1.235569	0.745581	0.379090	0.140870	-0.000995
	100	1.279259	0.977336	0.505194	0.218020	0.073045	0.007793
	200	1.060583	0.758660	0.364371	0.134859	0.028359	-0.003004
	500	0.770351	0.478421	0.169584	0.063237	0.030990	0.011171

O vício de Q_2, disposto na Tabela 4.4, também decresce quando o tamanho da amostra e o quantil aumentam, em todas as configurações de intervalo de censura. Os quantis são superestimados por Q_2 em todas as configurações, de modo geral.

Como no caso anterior, a magnitude do vício parece diferir entre as configurações, apresentado-se mais elevada para as configurações 2 e 4, nas quais há uma maior concentração de tempos de falha censurados à esquerda de U.

A magnitude do vício pode ser vista mais facilmente na Tabela 4.5, que apresenta o vício relativo de Q_2, os quais, de modo geral, decrescem quando o quantil e o

Tabela 4.3: *Vício relativo do estimador Q_1, segundo a configuração (C) e o tamanho da amostra (n).*

		Quantis					
C	n	0.01	0.02	0.05	0.10	0.20	0.50
1	50	0.7743	0.4901	0.1963	0.0801	0.0226	0.0027
	100	0.6122	0.3540	0.1184	0.0404	0.0153	-0.0003
	200	0.4860	0.2537	0.0756	0.0237	0.0067	0.0011
	500	0.3167	0.1403	0.0340	0.0122	0.0072	-0.0006
2	50	0.8700	0.5705	0.2495	0.1008	0.0212	-0.0020
	100	0.7058	0.4326	0.1544	0.0486	0.0171	-0.0015
	200	0.5813	0.3290	0.1123	0.0326	0.0060	0.0019
	500	0.4070	0.1940	0.0443	0.0081	0.0075	0.0007
3	50	0.7354	0.4574	0.1757	0.0693	0.0145	0.0035
	100	0.5866	0.3329	0.1062	0.0272	0.0102	0.0001
	200	0.4662	0.2372	0.0632	0.0167	0.0056	0.0022
	500	0.3008	0.1277	0.0209	0.0092	0.0065	0.0007
4	50	0.9712	0.6555	0.3133	0.1331	0.0410	-0.0002
	100	0.8081	0.5185	0.2123	0.0765	0.0213	0.0017
	200	0.6699	0.4025	0.1531	0.0473	0.0083	-0.0007
	500	0.4866	0.2538	0.0713	0.0222	0.0090	0.0024

tamanho da amostra aumentam.

Tabela 4.4: *Vício do estimador Q_2, segundo a configuração (C) e o tamanho da amostra (n).*

		Quantis					
C	n	0.01	0.02	0.05	0.10	0.20	0.50
1	50	0.427507	0.333930	0.190711	0.087140	0.006155	0.003886
	100	0.438302	0.318205	0.150785	0.051120	-0.00336	-0.00156
	200	0.425709	0.291171	0.118756	0.029581	-0.01045	0.008942
	500	0.334006	0.194792	0.054638	0.006339	0.00277	0.002456
2	50	0.712707	0.573885	0.358431	0.192666	0.059059	0.007910
	100	0.658455	0.496587	0.257576	0.105857	0.024746	0.001974
	200	0.618134	0.444842	0.213993	0.074376	0.005065	0.000155
	500	0.476240	0.294651	0.093509	0.015589	0.012785	0.003660
3	50	0.614839	0.472688	0.265742	0.123222	0.017423	0.010555
	100	0.561738	0.401415	0.183800	0.055572	0.003696	0.000513
	200	0.500581	0.330681	0.125559	0.034675	-0.00010	0.005598
	500	0.364625	0.202871	0.041859	0.011073	0.009082	0.003511
4	50	0.767750	0.634819	0.420854	0.248064	0.097700	0.007045
	100	0.737605	0.579611	0.338370	0.164441	0.049688	0.001332
	200	0.696697	0.524251	0.281590	0.118117	0.022685	0.002520
	500	0.559761	0.375541	0.150545	0.049929	0.013583	0.005413

O estimador Q_3 tem comportamento distinto entre as configurações. Na primeira, o seu vício decresce quando o tamanho da amostra e o quantil aumentam. Para as demais configurações, no entanto, o vício apresenta um comportamento irregular à medida que o tamanho da amostra aumenta, como se observa na Tabela 4.6.

Ainda para as configurações 2, 3 e 4, nota-se que o vício cresce inicialmente, em módulo, voltando a decrescer a partir de determinado quantil, para os tamanhos de amostra 50, 100 e 200 (exceto para a configuração 3, onde isto é válido para $n = 50$ e $n = 100$). Este ponto, a partir do qual o vício começa a decrescer, varia conforme

Tabela 4.5: *Vício relativo do estimador Q_2, segundo a configuração (C) e o tamanho da amostra (n).*

		Quantis					
C	n	0.01	0.02	0.05	0.10	0.20	0.50
1	50	0.2700	0.1771	0.0801	0.0306	0.0018	0.0009
	100	0.2769	0.1688	0.0634	0.0179	-0.0010	-0.0003
	200	0.2689	0.1545	0.0499	0.0104	-0.0030	0.0020
	500	0.2110	0.1033	0.0230	0.0022	0.0008	0.0005
2	50	0.4502	0.3044	0.1506	0.0676	0.0172	0.0017
	100	0.4159	0.2634	0.1082	0.0372	0.0072	0.0004
	200	0.3905	0.2360	0.0899	0.0261	0.0015	0.0000
	500	0.3008	0.1563	0.0393	0.0055	0.0037	0.0008
3	50	0.3884	0.2508	0.1117	0.0433	0.0051	0.0023
	100	0.3548	0.2129	0.0772	0.0195	0.0010	0.0001
	200	0.3162	0.1754	0.0528	0.0122	-0.0000	0.0012
	500	0.2303	0.1076	0.0176	0.0039	0.0027	0.0008
4	50	0.4846	0.3368	0.1769	0.0871	0.0284	0.0015
	100	0.4659	0.3075	0.1422	0.0577	0.0145	0.0003
	200	0.4401	0.2781	0.1183	0.0415	0.0066	0.0006
	500	0.3536	0.1992	0.0633	0.0175	0.0040	0.0012

a configuração e o tamanho da amostra. Mas, em geral, verifica-se que quanto maior o tamanho da amostra, tal ponto parece cada vez mais aproximar-se do primeiro quantil. Deste modo, o vício tende a ficar decrescente, em módulo, para amostras de maior tamanho, conforme se observa para $n = 500$ (exceto para a configuração 4).

No entanto, essa observação parece irrelevante no momento em que se verifica o vício relativo reduzir-se à medida que o quantil aumenta, como está exposto na Tabela 4.7.

Há diferença entre as configurações também no que concerne à magnitude do vício, que é consideravelmente maior para as configurações 2 e 4, como pode ser observado pela Tabela 4.7. Como nos estimadores anteriores, isto parece estar relacionado com a quantidade de tempos de falha censurados à esquerda de U.

De modo geral, o estimador Q_4 subestima os quantis em todas as configurações. Como revela a Tabela 4.8, o vício de Q_4, em geral, cresce quando o tamanho da amostra aumenta, para os quantis mais baixos ($q_1 = 0.01$ e $q_2 = 0.02$). Para os demais quantis, contudo, o vício diminui à medida que o tamanho da amostra aumenta.

É possível verificar ainda que o vício cresce inicialmente, ao aumentar o quantil. Entretanto, ele volta a decrescer a partir de determinado ponto. Nota-se que quanto maior o tamanho da amostra, o vício começa a decrescer cada vez mais próximo aos quantis iniciais.

O comportamento do vício relativo é bastante semelhante ao do vício.

Diferentemente dos demais estimadores, a magnitude do vício de Q_4 não apresenta grande diferença entre as configurações. E, notadamente, esta magnitude é menor para os quantis mais baixos (q_1 e q_2), quando comparada às obtidas nos casos anteriores.

Os vícios relativos, constantes das Tabelas 4.3, 4.5, 4.7 e 4.9, mostram que todos os estimadores considerados apresentam vício muito pequeno para a mediana ($q_6 = 0.50$). Para os quantis mais baixos, o estimador Q_4 é o que proporciona o menor vício. Para os demais quantis, os vícios são menores para o estimador Q_2.

Tabela 4.6: *Vício do estimador Q_3, segundo a configuração (C) e o tamanho da amostra (n).*

		Quantis					
C	n	0.01	0.02	0.05	0.10	0.20	0.50
1	50	0.691998	0.528810	0.278966	0.130465	0.032068	0.010219
	100	0.636109	0.449009	0.202519	0.073073	0.024301	-0.003812
	200	0.550143	0.359335	0.137638	0.046136	0.002241	0.005569
	500	0.398612	0.220104	0.068088	0.016913	0.017386	0.000466
2	50	-1.33101	-1.48714	-1.53685	-1.29041	-0.88586	0.026192
	100	-1.38662	-1.53692	-1.51626	-1.20903	-0.80749	0.000852
	200	-1.47302	-1.65869	-1.58295	-1.24177	-0.82455	-0.007189
	500	-0.59234	-0.52068	-0.42739	-0.34702	-0.24257	-0.009921
3	50	-0.76882	-0.81111	-0.77229	-0.65540	-0.46589	-0.001892
	100	-0.79889	-0.80949	-0.73063	-0.59893	-0.41557	-0.025629
	200	-0.91607	-0.89085	-0.74486	-0.59628	-0.40631	-0.019099
	500	-1.02257	-0.93165	-0.74678	-0.58689	-0.39691	-0.019861
4	50	-1.47329	-1.68828	-1.84273	-1.65030	-1.21402	0.0410959
	100	-1.54094	-1.78530	-1.98080	-1.66406	-1.13572	0.0162171
	200	-1.56950	-1.84584	-2.09267	-1.68920	-1.12455	0.0088513
	500	-1.58239	-1.87934	-2.22159	-1.75526	-1.15749	-0.006326

Tabela 4.7: *Vício relativo do estimador Q_3, segundo a configuração (C) e o tamanho da amostra (n).*

		Quantis					
C	n	0.01	0.02	0.05	0.10	0.20	0.50
1	50	0.4371	0.2805	0.1172	0.0458	0.0093	0.0022
	100	0.4018	0.2382	0.0851	0.0257	0.0071	-0.0008
	200	0.3475	0.1906	0.0578	0.0162	0.0007	0.0012
	500	0.2518	0.1168	0.0286	0.0059	0.0051	0.0001
2	50	-0.8408	-0.7889	-0.6459	-0.4530	-0.2578	0.0057
	100	-0.8759	-0.8153	-0.6372	-0.4244	-0.2350	0.0002
	200	-0.9305	-0.8799	-0.6652	-0.4359	-0.2399	-0.0016
	500	-0.3742	-0.2762	-0.1796	-0.1218	-0.0706	-0.0022
3	50	-0.4856	-0.4303	-0.3246	-0.2301	-0.1356	-0.0004
	100	-0.5046	-0.4294	-0.3071	-0.2103	-0.1209	-0.0056
	200	-0.5786	-0.4726	-0.3130	-0.2093	-0.1182	-0.0042
	500	-0.6459	-0.4942	-0.3138	-0.2060	-0.1155	-0.0044
4	50	-0.9306	-0.8956	-0.7744	-0.5793	-0.3533	0.0090
	100	-0.9734	-0.9471	-0.8324	-0.5842	-0.3305	0.0036
	200	-0.9914	-0.9792	-0.8795	-0.5930	-0.3272	0.0019
	500	-0.9995	-0.9970	-0.9336	-0.6162	-0.3368	-0.0014

Tabela 4.8: *Vício do estimador Q_4, segundo a configuração (C) e o tamanho da amostra (n).*

		Quantis					
C	n	0.01	0.02	0.05	0.10	0.20	0.50
1	50	0.025925	-0.16170	-0.34558	-0.35719	-0.21534	-0.04753
	100	-0.05995	-0.22377	-0.34446	-0.28091	-0.10567	-0.02342
	200	-0.09584	-0.22298	-0.28089	-0.17857	-0.04713	0.00412
	500	-0.14979	-0.22166	-0.19295	-0.07711	-0.00253	-0.00190
2	50	-0.01797	-0.18810	-0.34879	-0.34797	-0.20469	-0.04895
	100	-0.08277	-0.23185	-0.34050	-0.26804	-0.09626	-0.02170
	200	-0.10452	-0.22458	-0.26729	-0.15517	-0.04213	-0.01233
	500	-0.18470	-0.27572	-0.23366	-0.08499	-0.00860	0.00124
3	50	-0.04268	-0.18559	-0.30723	-0.27742	-0.14494	-0.03214
	100	-0.10192	-0.21535	-0.27652	-0.19039	-0.05694	-0.01260
	200	-0.11304	-0.18670	-0.19005	-0.08757	-0.02292	-0.00173
	500	-0.14552	-0.18818	-0.12932	-0.02951	0.005029	-0.00024
4	50	-0.02867	-0.19307	-0.34795	-0.35194	-0.22315	-0.08542
	100	-0.07786	-0.21572	-0.31394	-0.26201	-0.09879	-0.03862
	200	-0.05653	-0.15159	-0.18029	-0.10204	-0.03071	-0.01576
	500	-0.07811	-0.12363	-0.10478	-0.02608	-0.00017	0.00242

Tabela 4.9: *Vício relativo do estimador Q_4, segundo a configuração (C) e o tamanho da amostra (n).*

		Quantis					
C	n	0.01	0.02	0.05	0.10	0.20	0.50
1	50	0.0164	-0.0858	-0.1452	-0.1254	-0.0627	-0.0104
	100	-0.0379	-0.1187	-0.1448	-0.0986	-0.0307	-0.0051
	200	-0.0605	-0.1183	-0.1180	-0.0627	-0.0137	0.0009
	500	-0.0946	-0.1176	-0.0811	-0.0271	-0.0007	-0.0004
2	50	-0.0114	-0.0998	-0.1466	-0.1222	-0.0596	-0.0107
	100	-0.0523	-0.1230	-0.1431	-0.0941	-0.0280	-0.0048
	200	-0.0660	-0.1191	-0.1123	-0.0545	-0.0123	-0.0027
	500	-0.1167	-0.1463	-0.0982	-0.0298	-0.0025	0.0003
3	50	-0.0270	-0.0985	-0.1291	-0.0974	-0.0422	-0.0070
	100	-0.0644	-0.1142	-0.1162	-0.0668	-0.0166	-0.0028
	200	-0.0714	-0.0990	-0.0799	-0.0307	-0.0067	-0.0004
	500	-0.0919	-0.0998	-0.0543	-0.0104	0.0015	-0.0001
4	50	-0.0181	-0.1024	-0.1462	-0.1235	-0.0649	-0.0187
	100	-0.0492	-0.1144	-0.1319	-0.0920	-0.0287	-0.0085
	200	-0.0357	-0.0804	-0.0758	-0.0358	-0.0089	-0.0035
	500	-0.0493	-0.0656	-0.0440	-0.0092	-0.0001	0.0005

4.1.2 Variância dos Estimadores

De forma geral, para Q_1, a variância decresce quando aumenta o tamanho da amostra. Mas, em relação aos quantis, permanece praticamente constante até o quantil $q_4 = 0.10$, a partir de quando começa a decrescer. Para amostras de tamanho 500, todavia, a variância começa a decrescer a partir do quantil $q_3 = 0.05$, como se vê na Tabela 4.10.

Tabela 4.10: *Variância do estimador Q_1, segundo a configuração (C) e o tamanho da amostra (n).*

		Quantis					
C	n	0.01	0.02	0.05	0.10	0.20	0.50
1	50	0.217213	0.217213	0.227778	0.224348	0.209240	0.180025
	100	0.161281	0.161281	0.173256	0.163995	0.122162	0.097482
	200	0.128398	0.124144	0.135191	0.111343	0.083410	0.066344
	500	0.098887	0.103027	0.083331	0.065230	0.043718	0.036416
2	50	0.209128	0.209128	0.212795	0.222811	0.216132	0.169717
	100	0.166902	0.166902	0.175339	0.177643	0.141107	0.097785
	200	0.135881	0.135678	0.135444	0.126454	0.096726	0.052957
	500	0.111167	0.110940	0.106124	0.083913	0.043917	0.027265
3	50	0.190019	0.190019	0.200737	0.213486	0.198560	0.176893
	100	0.146542	0.147011	0.162278	0.160450	0.118716	0.090354
	200	0.120163	0.117526	0.121991	0.108529	0.070515	0.050351
	500	0.092106	0.096283	0.085024	0.054344	0.033132	0.026569
4	50	0.230596	0.230596	0.230510	0.249870	0.246898	0.185406
	100	0.191113	0.191113	0.190176	0.196800	0.167775	0.104093
	200	0.153378	0.153378	0.161416	0.156348	0.115160	0.062208
	500	0.131329	0.130385	0.129689	0.092124	0.058584	0.030309

Na tabela 4.11 é possível ver que, para Q_2, a variância decresce à medida que aumentam o tamanho da amostra e o quantil, independentemente da configuração de intervalo de censura. Entre os estimadores considerados neste trabalho, este é o único

que apresenta esta característica.

Tabela 4.11: *Variância do estimador Q_2, segundo a configuração (C) e o tamanho da amostra (n).*

		Quantis					
C	n	0.01	0.02	0.05	0.10	0.20	0.50
1	50	0.171554	0.158255	0.133142	0.110465	0.086978	0.075587
	100	0.127962	0.117801	0.094780	0.070092	0.050425	0.039018
	200	0.097361	0.085614	0.067687	0.049456	0.034121	0.024879
	500	0.074491	0.064888	0.042271	0.027872	0.016655	0.012998
2	50	0.188543	0.177544	0.157243	0.137559	0.110263	0.074589
	100	0.149798	0.143288	0.126406	0.097754	0.068542	0.041968
	200	0.113133	0.105907	0.090571	0.069638	0.047245	0.021900
	500	0.094458	0.087954	0.067176	0.044422	0.020006	0.010188
3	50	0.169922	0.159709	0.138765	0.119912	0.097309	0.082529
	100	0.127045	0.120367	0.104024	0.084101	0.053431	0.042022
	200	0.099345	0.089902	0.073487	0.056073	0.033051	0.021294
	500	0.075234	0.068601	0.048700	0.028222	0.015294	0.011455
4	50	0.213587	0.201963	0.181142	0.157572	0.129014	0.077007
	100	0.167927	0.160418	0.141990	0.117846	0.082356	0.045046
	200	0.130391	0.124178	0.108573	0.085635	0.055915	0.025372
	500	0.110976	0.103104	0.081541	0.050173	0.027057	0.011904

Pela Tabela 4.12, verifica-se que a variância de Q_3 decresce quando aumentam o tamanho da amostra e o quantil, na configuração 1.

Para as demais configurações, a variância decresce à medida que o tamanho da amostra aumenta. Contudo, quando observado o comportamento da variância em relação ao quantil, nota-se que a variância cresce inicialmente e volta a decrescer a partir de determinado quantil. O ponto a partir do qual se inicia o decréscimo da variância muda de acordo com a configuração e o tamanho da amostra. Em todos os casos, entretanto, verifica-se que este ponto parece aproxima-se dos quantis iniciais

quando o tamanho da amostra aumenta.

Tabela 4.12: *Variância do estimador Q_3, segundo a configuração (C) e o tamanho da amostra (n).*

		Quantis					
C	n	0.01	0.02	0.05	0.10	0.20	0.50
1	50	0.221069	0.182317	0.146009	0.134122	0.114363	0.101713
	100	0.134754	0.119984	0.105329	0.091024	0.073746	0.053326
	200	0.098219	0.087901	0.084053	0.068310	0.047848	0.035995
	500	0.075481	0.072881	0.055494	0.038696	0.024543	0.019319
2	50	0.230398	0.371282	0.595480	0.562709	0.370357	0.058347
	100	0.145454	0.268275	0.447770	0.317223	0.154436	0.025951
	200	0.069016	0.162180	0.303957	0.198553	0.092058	0.013909
	500	0.042826	0.035433	0.023779	0.015059	0.008832	0.005324
3	50	0.571781	0.625194	0.557401	0.375432	0.173131	0.069260
	100	0.421990	0.461489	0.380569	0.212778	0.089370	0.025729
	200	0.272369	0.314896	0.217514	0.117523	0.047526	0.013031
	500	0.115953	0.138918	0.084503	0.047598	0.019657	0.005270
4	50	0.122340	0.211382	0.439933	0.640084	0.682640	0.055162
	100	0.038261	0.085777	0.267882	0.351763	0.278771	0.025874
	200	0.012476	0.030038	0.155041	0.176027	0.122295	0.014687
	500	0.000115	0.001628	0.062392	0.054735	0.030487	0.005812

A variância de Q_4 decresce quando aumenta o tamanho da amostra para todas as configurações, como mostra a Tabela 4.13.

Ao observar-se o comportamento da variância de Q_4 em relação ao quantil, nota-se um crescimento inicial. No entanto, a variância volta a decrescer a partir de determinado ponto, o qual varia de acordo com a configuração e o tamanho da amostra. Este é o mesmo comportamento apresentado pela variância de Q_3 para as configurações 1, 2 e 3. A diferença é que, para a variância de Q_4, isto se verifica para todas as configurações.

Ainda observando a Tabela 4.13, verifica-se que a magnitude da variância na configuração 1 nos dois primeiros quantis é bastante inferior àquelas obtidas nas demais configurações.

Tabela 4.13: *Variância do estimador Q_4, segundo a configuração (C) e o tamanho da amostra (n).*

		Quantis					
C	n	0.01	0.02	0.05	0.10	0.20	0.50
1	50	0.082475	0.099886	0.127748	0.117062	0.092028	0.10038
	100	0.066679	0.080068	0.092817	0.065921	0.050114	0.049979
	200	0.049313	0.067807	0.073267	0.047154	0.034384	0.033608
	500	0.041994	0.056735	0.047367	0.025403	0.018953	0.018127
2	50	0.196531	0.243400	0.304926	0.314666	0.233165	0.095576
	100	0.146587	0.185667	0.217230	0.187567	0.081528	0.051617
	200	0.106401	0.136093	0.137455	0.107918	0.054917	0.02562
	500	0.073100	0.086705	0.075926	0.040733	0.018777	0.012685
3	50	0.184788	0.238462	0.277993	0.245891	0.154886	0.092318
	100	0.157389	0.207127	0.231282	0.148296	0.073985	0.050183
	200	0.134245	0.172410	0.147516	0.071443	0.039745	0.027142
	500	0.105723	0.112059	0.059649	0.028137	0.018559	0.015177
4	50	0.312288	0.388977	0.511883	0.524592	0.378491	0.10445
	100	0.253366	0.321061	0.407482	0.378871	0.170763	0.056578
	200	0.192518	0.254376	0.2755	0.199998	0.088808	0.033504
	500	0.183244	0.227931	0.183234	0.079991	0.035311	0.017664

De modo geral, constata-se que a variância do estimador Q_2 apresenta-se menor que a dos demais. Excetuam-se os dois primeiros quantis analisados na configuração 4, para os quais Q_3 apresenta a menor variância, e os dois primeiros quantis na configuração 1, para os quais Q_4 apresenta a menor variância. Para os demais quantis da configuração 1, Q_4 tem variância semelhante a de Q_2.

A apresentação gráfica da variância dos estimadores encontra-se no Apêndice 1,

representadas pelas Figuras 5.5, 5.6, 5.7 e 5.8.

4.1.3 Outras Considerações

A Figura 4.1 ilustra a distribuição da função quantílica de uma amostra de tamanho 100 e as estimativas obtidas pelo emprego dos quatro estimadores sob consideração, segundo a configuração de intervalo de censura.

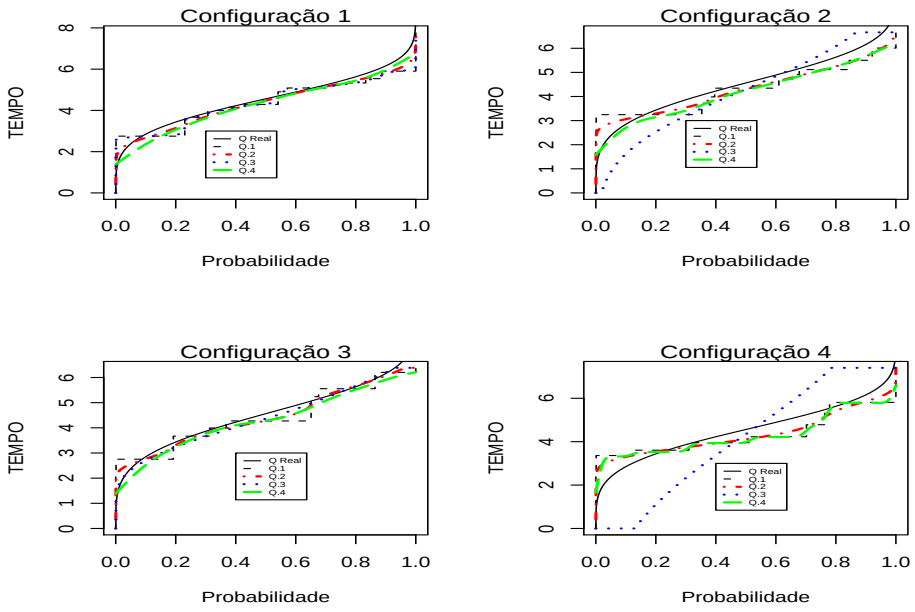


Figura 4.1: *Comparação entre os estimadores da função quantílica.*

A grande diferença verificada entre os estimadores Q_2 e Q_3 é devida unicamente ao método de obtenção do parâmetro de suavização.

Para as configurações 2, 3 e 4, a janela obtida pelo método de Pan, em média, é bem superior àquela encontrada pelo método da validação cruzada modificado. Em consequência, ocorre uma supersuavização do ENPMV $\hat{F}$ que, ao ser invertido, proporciona estimativas ruins para a função quantílica.

Ainda foram elaborados os histogramas das estimativas dos quantis, para todos os estimadores e todas as configurações, com a finalidade de observar seu comportamento.

Os histogramas estão contidos no Apêndice 2 e revelam um comportamento semelhante para a distribuição das estimativas produzidas por Q_1 e por Q_2. Em todas

as configurações, ambos apresentaram assimetria à esquerda para os quantis q_1 a q_4, parecendo ter distribuição normal para as estimativas de q_5 e q_6.

O comportamento das distribuições das estimativas obtidas empregando Q_3 é semelhante ao descrito anteriormente apenas para a configuração 1. Para as demais configurações, somente a mediana parece apresentar uma distribuição próxima da Normal.

Já a distribuição das estimativas produzidas por Q_4 apresentam assimetria à direita.

4.2 Aplicação

Os estimadores considerados na dissertação foram aplicados também a um conjunto de dados reais, os quais estão descritos em Finkelstein and Wolfe (1985). Estes dados referem-se ao tempo (em meses) até a deterioração cosmética do tecido para pacientes em tratamento de câncer de mama, submetidas a radioterapia.

As 94 observações originais contêm apenas o intervalo de tempo onde ocorreu a deterioração, sendo que o estudo se encerrou aos 61 meses de observação. Ou seja, ocorre censura à direita em V quando o limite superior de observação é igual a 61 meses. Os autores ainda descrevem no artigo que, quando o limite inferior de observação é igual a 0 meses, houve censura à esquerda em U.

A Tabela 4.14 traz as estimativas obtidas para os quantis considerados neste trabalho, segundo o estimador empregado.

Tabela 4.14: *Estimativas (em meses) resultantes da aplicação a dados reais, segundo o estimador.*

	Quantis					
Estimador	0.01	0.02	0.05	0.10	0.20	0.50
Q_1	5	5	5	8	11	19
Q_2	1.113686	2.566643	4.998787	7.591026	11.93035	23.69735
Q_3	2.419579	3.562584	5.555397	7.774391	11.81307	22.72002
Q_4	3.000441	3.661415	5.095275	7.992579	11.03038	20.93262

A configuração de intervalos de censura observada nos dados reais não tem exatamente o mesmo padrão das configurações consideradas na simulação. Para os dados reais, constata-se que apenas 5% dos dados foram censurados à esquerda, 54% foram censurados no intervalo (U, V) e 41%, à direita de V. É possível que o comportamento dos estimadores para estes dados se assemelhe a um dos comportamentos observados na simulação, embora não se possa afirmar qual.

Todavia, nota-se que as maiores diferenças percebidas entre as estimativas estão nos primeiros quantis, como era de se esperar.

A Figura 4.2 mostra as funções quantílicas estimadas, por cada um dos estimadores estudados.

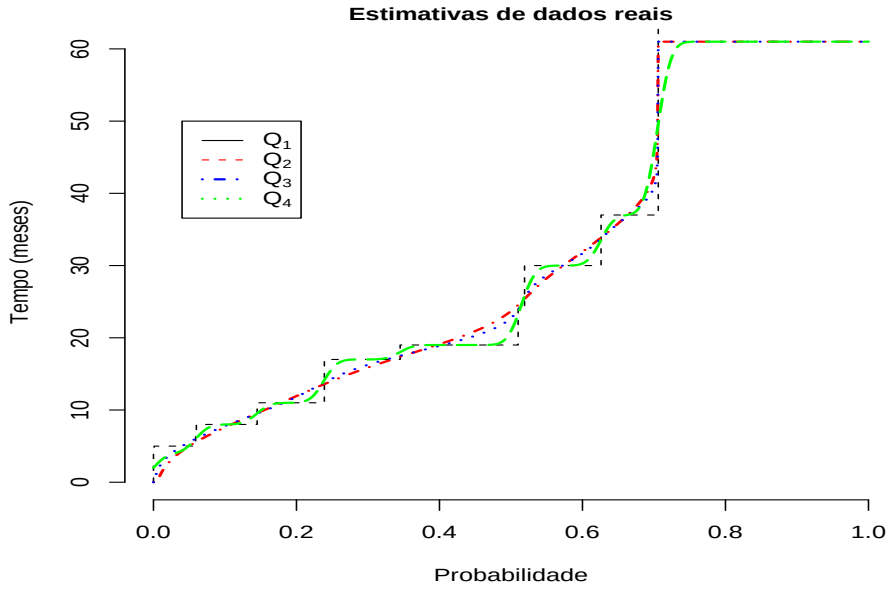


Figura 4.2: *Comparação entre os estimadores da função quantílica, para dados reais.*

Capítulo 5

Conclusões e Trabalhos Futuros

5.1 Conclusões

Os resultados descritos no Capítulo anterior permitem algumas conclusões acerca da variância e do vício dos estimadores da função quantílica, as quais serão apresentadas a seguir.

A variância dos estimadores diminui quando o tamanho da amostra aumenta, fixado o quantil, o que era esperado. Entretanto, somente a variância de Q_2 diminui quando o quantil aumenta. As variâncias dos estimadores Q_3 e Q_4 apresentam diferenças significativas entre as configurações de intervalo de censura, o que não ocorre para Q_1 e Q_2.

O vício relativo diminui quando o tamanho da amostra aumenta apenas para Q_1 e Q_2. Ao aumentar o quantil, diminui o vício relativo dos estimadores Q_1, Q_2 e Q_3. A magnitude do vício parece estar relacionada com a censura à esquerda de U, especialmente para os quantis mais baixos, tornando-se maior o vício quanto maior o percentual de censura.

Em relação aos quantis mais baixos, ressalta-se que a magnitude da variância de Q_4 é a menor observada para os dois primeiros quantis da configuração 1. Analisando os resultados para as configurações 2 e 3, vê-se que a variância de Q_4 está próxima à variância de Q_2. Para a configuração 4, contudo, a variância de Q_4 é superior à de Q_2.

Já o vício relativo de Q_4, para os dois primeiros quantis, é muito inferior ao

observado para os demais estimadores.

No que concerne a estimativa da mediana, todos os estimadores apresentam vício pequeno. Contudo, em média, a estimativa de Q_2 aproxima-se mais do valor real do quantil. Além disto, este estimador apresenta menor variabilidade, notadamente para amostras de menor tamanho.

Destaca-se, ainda, que Q_3 apresenta comportamento bastante distinto entre as configurações de intervalo de censura, mostrando-se muito instável. Ademais, o emprego deste estimador exige um esforço computacional muito grande, em função do cálculo da janela ótima, para o qual a simulação mostra não haver resultados recompensadores. Por depender da definição da partição da amostra, para obter a janela ótima, este estimador necessita de estudos mais detalhados.

Finalmente, o estimador Q_4 parece ser o mais adequado para estimar q_1 e q_2, enquanto Q_2 apresenta comportamento mais estável, parecendo mais apropriado para estimar os demais quantis considerados nesta dissertação.

5.2 Trabalhos Futuros

Os estudos de simulação mostraram ser provável uma relação entre o parâmetro de suavização e as estimativas da função quantílica. Assim, sugere-se, para trabalhos futuros, o emprego de outros métodos de seleção da janela ótima.

Sugere-se, ainda, considerar uma variação do estimador Q_2, obtido pela suavização do ENPMV $\hat{F}$, utilizando o método do núcleo com uma janela variável. O objetivo é melhorar a suavização e verificar sua influência sobre a estimação da função quantílica.

Nas Subseções seguintes são apresentados os métodos Bayesiano e Bootstrap para selecionar o parâmetro de suavização. Da mesma forma, descreve-se o estimador da função quantílica, considerando a janela variável na suavização do ENPMV.

5.2.1 Método Bayesiano

Outra abordagem para selecionar o parâmetro de suavização é o emprego do critério de Bayes, proposto por Kulasekera e Padgett (2006). A grande vantagem deste método reside no fato de que a janela Bayesiana é exata para todos os tamanhos de amostra. Os autores sugerem o emprego de uma função núcleo assimétrica, cujo uso elimina a possibilidade de a função densidade estimada apresentar valores negativos.

Para uma *priori* de h, $\xi(h)$, a distribuição à *posteriori* é dada por

$$\xi(h|x) = \frac{f_h(x)\xi(h)}{\int f_h(x)\xi(h)dh}.$$

Como, em geral, não se conhece a função densidade $f_h(x)$, os autores propõem estimar a *posteriori* substituindo $f_h(x)$ por sua estimativa. Dessa forma, no caso de censura intervalar, dadas as realizações de $\tau = \{(u_i, v_i, \delta_i, \gamma_i),\ i = 1, \ldots, n\}$, a *posteriori* pode ser estimada por

$$\hat{\xi}(h|x,\tau) = \frac{\hat{f}_h(x)\xi(h)}{\int \hat{f}_h(x)\xi(h)dh}.$$

Assim, é possível estimar o parâmetro de suavização pela média da *posteriori*

$$\tilde{h}(x) = \int h\hat{\xi}(h|x,\tau)dh. \tag{5.1}$$

Com esta abordagem, a densidade a *posteriori* é uma função somente de h e os valores de $\hat{\xi}$ e $\tilde{h}$ podem, em alguns casos, ser obtidos explicitamente, a partir de uma estrutura simples de *priori*.

A janela h obtida por (5.1) não converge necessariamente para zero quando $n \to \infty$. Contudo, segundo Kulasekera e Padgett (2006), é possível escolher uma *priori* que atenda a esta condição.

Os autores ainda compararam os resultados da escolha de h empregando o método Bayesiano e o método de validação cruzada, mediante a razão entre o Erro Quadrático Médio Estimado (EQME) das duas abordagens, para dados censurados à direita. Considere $EQME(\phi_n(t)) = \sum_{i=1}^{N}(\phi_n(t) - \phi(t))^2/N$ para um dado t, uma função ϕ e um número N de simulações. Em todos os casos, o estimador da densidade pelo critério de Bayes foi superior para valores pequenos e moderados de t.

5.2.2 Seleção da Janela pelo Método Bootstrap

Além dos métodos já citados para obtenção do parâmetro de suavização ótimo, para o estimador da função quantílica apresentado em (3.23), também se quer destacar o cálculo de h pelo método bootstrap.

Este método consiste na retirada de B novas amostras aleatórias de tamanho m, com reposição, a partir da amostra original de tamanho n, com probabilidade $1/n$ de uma observação ser selecionada.

Defina o estimador de $Q(p)$ baseado nas amostras bootstrap por

$$\hat{Q}_n^B(p) = \inf\{t : \hat{F}_n^B(t) \geq p\}, \quad 0 \leq p \leq 1,$$

onde $\hat{F}_n^B(t) = \frac{1}{B}\sum_{j=1}^{B}\hat{F}_{n,j}(t)$, sendo $\hat{F}_{n,j}$ a estimativa (ENPMV) de F obtida utilizando a j-ésima amostra bootstrap.

Defina ainda o estimador de $Q(p)$ a partir da função núcleo, também baseado nas amostras bootstrap como

$$\begin{aligned}\check{Q}_n^B(p) &= \int_0^1 K\left(\frac{t-p}{h}\right) d\hat{Q}_n^B(p) \\ &= \sum_{i=1}^{n} K\left(\frac{t-p}{h}\right)\left[\hat{Q}_n^B(p_{j+1}) - \hat{Q}_n^B(p_j)\right].\end{aligned}$$

Então, pode-se obter o estimador da variância de $\check{Q}_n(p)$ por

$$\hat{\mathrm{Var}}(\check{Q}_n(p)) = \frac{1}{B-1}\sum_{i=1}^{B}\left(\check{Q}_{n,i}^B(p) - \frac{\sum_{i=1}^{B}\check{Q}_{n,i}^B(p)}{B}\right)^2$$

e o estimador do vício de $\check{Q}_n(p)$ por

$$\hat{b}_1(\check{Q}_n(p)) = \frac{1}{B}\sum_{i=1}^{B}\check{Q}^B_{n,i}(p) - \hat{Q}_n(p).$$

Seja $EQM^B_{\check{Q}_n(p)}$ o erro quadrático médio do estimador bootstrap de $\check{Q}_n(p)$. Padgett e Thombs (1986) analisaram o $EQM_{\check{Q}_n(p)}$ e, mediante simulações para dados censurados à direita, observaram que, fixando o percentil p e aumentando o valor de h, o vício aumentava enquanto a variância diminuía. Desta forma, é possível que $EQM_{\check{Q}_n(p)}$, como função de h, seja inicialmente decrescente e depois crescente.

Assim, a estimativa $EQM^B_{\check{Q}_n(p)}$ forneceria o valor de h que minimizaria o $EQM_{\check{Q}_n(p)}$. Contudo, Balbino, C. A. S. (2006) descreve situações com $EQM^B_{\check{Q}_n(p)}$ estritamente decrescente, devido a supersuavização de $\check{Q}^B_n(p)$. Para contornar o problema, o autor sugere trocar o estimador quantílico do núcleo sem bootstrap, $\check{Q}_n(p)$, pelo estimador quantílico obtido a partir do ENPMV de F, $\hat{Q}_n(p)$, no cálculo do vício, obtendo

$$\hat{b}_2(\check{Q}_n(p)) = \frac{1}{B}\sum_{i=1}^{B}\check{Q}^B_{n,i}(p) - \hat{Q}_n(p).$$

5.2.3 Outro Estimador da Função Quantílica

Como visto anteriormente, pode-se estimar a função quantílica mediante a inversão do ENPMV $\hat{F}$ de F, suavizado pelo método do núcleo estimador, obtendo

$$\tilde{Q}_n(p) = \inf\{t : \tilde{F}_n(t) \geq p\}, \quad 0 \leq p \leq 1,$$

onde,

$$\begin{aligned}\tilde{F}_n(t) &= \int_{-\infty}^{\infty}\mathcal{K}\left(\frac{t-x}{h}\right)d\hat{F}_n(x)\\ &= \sum_{j=1}^{m}\mathcal{K}\left(\frac{t-M_j}{h}\right)\left(\hat{F}_n(M_j) - \hat{F}_n(M_{j-1})\right),\end{aligned}$$

sendo $\mathcal{K}$ uma função de distribuição.

Uma variação desta estimação pode ser obtida mediante a suavização do ENPMV de F, $\hat{F}$, utilizando um parâmetro de suavização, h, variável. Assim, após inverter $\hat{F}$ obtém-se

$$\tilde{Q}'_n(p) = \inf\{t : \tilde{F}'_n(t) \geq p\}, \quad 0 \leq p \leq 1,$$

onde,

$$\begin{aligned}\tilde{F}'_n(t) &= \int_{-\infty}^{\infty} \mathcal{K}\left(\frac{t-x}{h(x)}\right) d\hat{F}_n(x) \\ &= \sum_{i=1}^{n} \mathcal{K}\left(\frac{t-M_i}{h(M_i)}\right)\left(\hat{F}_n(M_i) - \hat{F}_n(M_{i-1})\right).\end{aligned}$$

Neste caso, a janela ótima $h(X_i)$ varia de acordo com o intervalo entre os saltos da função de distribuição estimada e pode ser calculada empregando qualquer um dos métodos discutidos anteriormente.

Referências Bibliográficas

[1] Barlow, R. E., Bartholomew, D. J., Bremner, J. M. e Brunk, H. D. *Statistical Inference Under Order Restrictions*. New York: John Wiley & Sons, 1972.

[2] Balbino, C. A. S. *Núcleo Estimador para Função de Densidade e Função Quantílica na Presença de Censura à Direita*. Belo Horizonte, MG: Dissertação de Mestrado, Departamento de Estatística, UFMG, 2006.

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

[4] Colosimo, E. A. e Giolo, S. R. *Análise de Sobrevivência Aplicada*, São Paulo: Edgar Blücher, 2006.

[5] Finkelstein, D. M. e Wolfe, R. A. "A semi-parametric model for regression analysis of interval censored failure time data", *Biometrics*, **41**, 933-945, 1985.

[6] Gomes, A. E. 2006. *Implementação do algoritmo ICM na linguagem R*. Brasília: Universidade de Brasília.

[7] Groeneboom, P. e Wellner, J.A. *Information Bounds and Nonparametric Maximum Likelihood Estimation*, Birkhauser Verlag, 1992.

[8] Hoel, D.G., Walburg, H.E., Jr. "Statistical analysis of survival experiments", *Journal of the National Cancer Institute*, **49**, 361-362, 1972.

[9] Kulasekera, K.B. e Padgett, W.J. "Bayes bandwidth selection in kernel density estimation with censored data", *Journal of Nonparametric Statistics*, **18(2)**, 129-144, 2006.

[10] Lindsey, J. C. e Ryan, L. M. "Tutorial in Biostatistics Methods for Interval-Censored Data", *Statistics in Medicine*, **17**, 219-238, 1998.

[11] Marron, J. S. e Padgett, W. J. "Asymptotically Optimal Bandwidth Selection for Kernel Density Estimators from Randomly Right-Censored Samples", *The Annals of Statistics*, **15**, 1520-1535, 1987.

[12] Padgett, W. J. "A Kernel-type Estimator of a Quantile Function from Right-Censored Data", *Journal of the American Statistical Association*, **81**, 215-222, 1986.

[13] Padgett, W. J. e Thombs, L. A. "Smooth Nonparametric Quantile Estimation Under Censoring: Simulation and Bootstrap Methods", *Communications in Statistics - Simulation and Computation*, **15(4)**, 1003-1025, 1986.

[14] Pan, W. "Smooth estimation of the survival function for interval censored data", *Statistics in Medicine*, **19**, 2611-2624, 2000.

[15] Parzen, E. "On Estimation of a Probability Density Function and Mode", *Annals of Mathematical Statistics*, **33**, 1065-1076, 1962.

[16] Rosenblatt, M. "Remarks on Some Nonparametric Estimates of a Density Function", *Annals of Mathematical Statistics*, **27**, 832-837, 1956.

[17] Stone, M. "An asymptotically optimal window selection rule for kernel density estimates", *The Annals of Statistics*, **12**, 1285-1297, 1984.

[18] Silverman, B. W. *Density Estimation for Statistics and Data Analysis*, New York: Chapman and Hall, 1986.

Apêndice 1

Gráficos da variância e do valor absoluto do vício relativo.

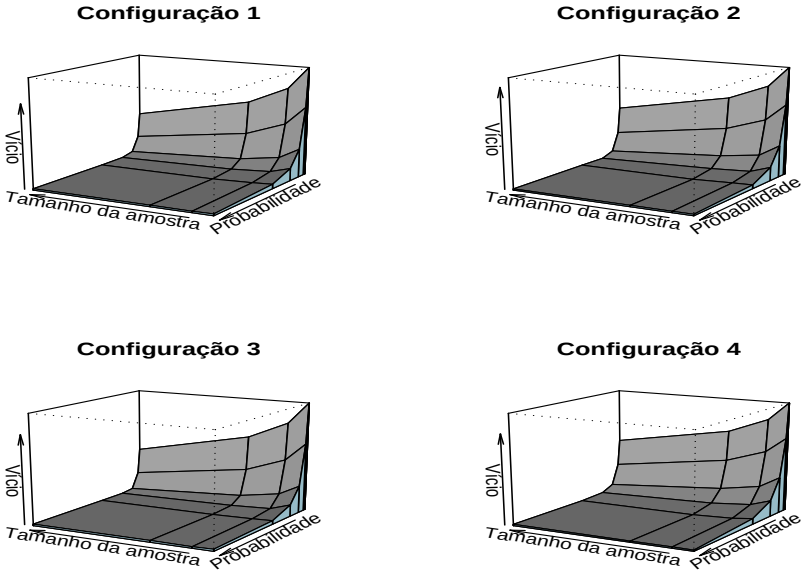
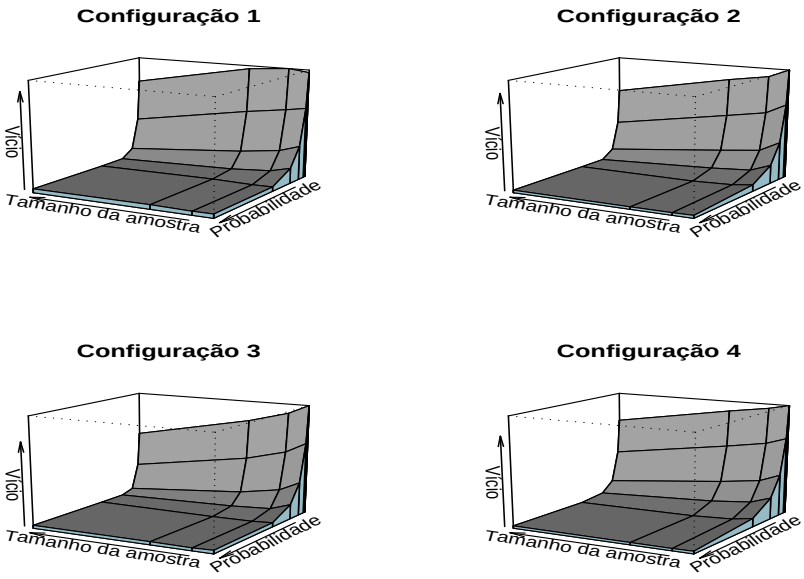
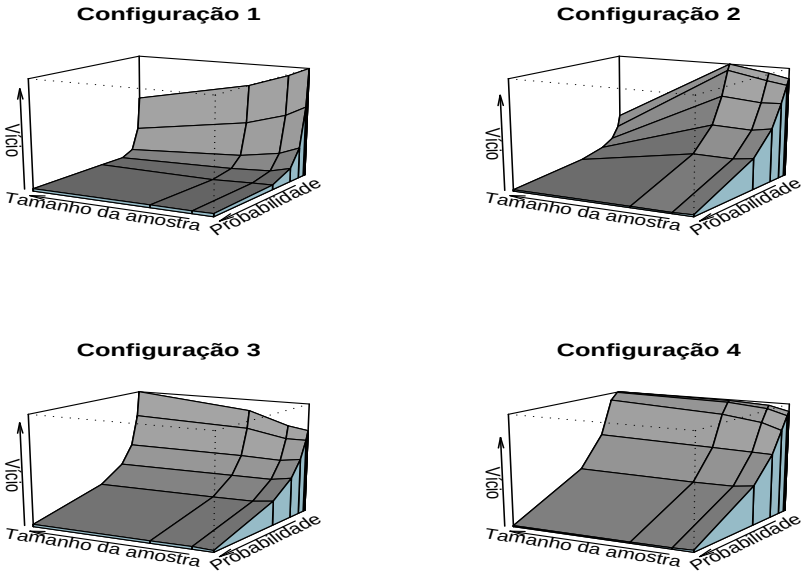


Figura 5.1: *Vício relativo do estimador Q_1, em módulo.*



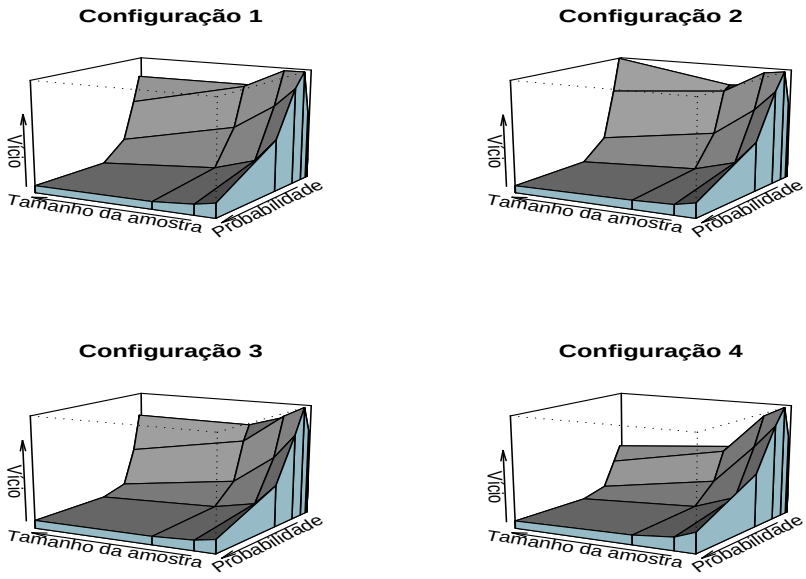
Vício relativo de Q.2, segundo o quantil e o tamanho da amostra, por configuração.

Figura 5.2: *Vício relativo do estimador Q_2, em módulo.*



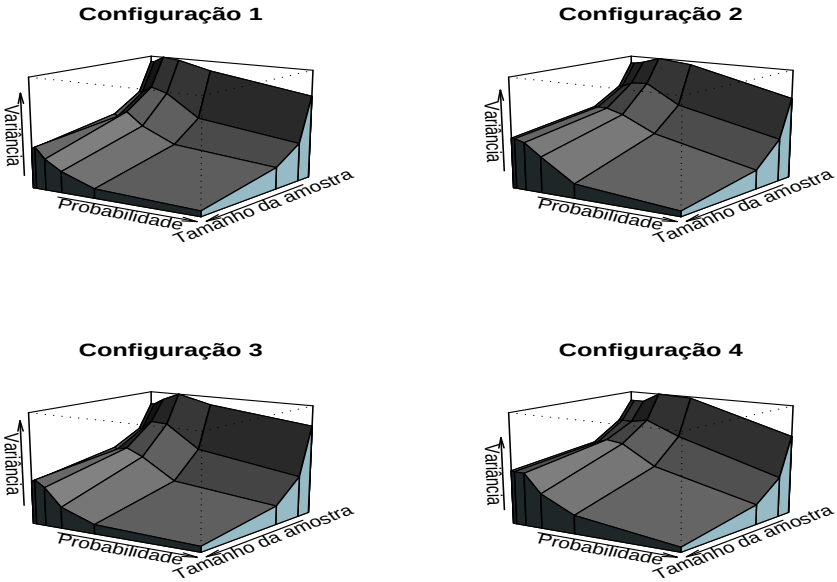
Vício relativo de Q.3, segundo o quantil e o tamanho da amostra, por configuração.

Figura 5.3: *Vício relativo do estimador Q_3, em módulo.*



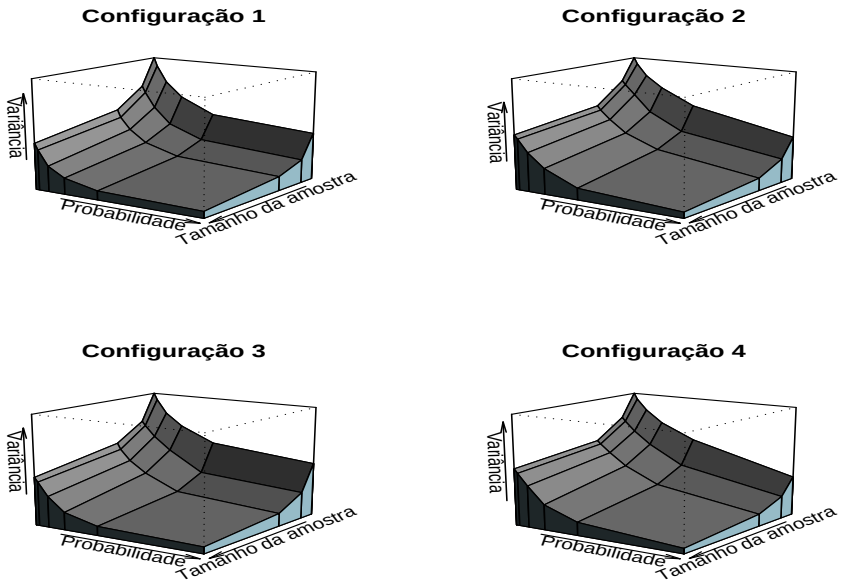
Vício relativo de Q.4, segundo o quantil e o tamanho da amostra, por configuração.

Figura 5.4: *Vício relativo do estimador Q_4, em módulo.*



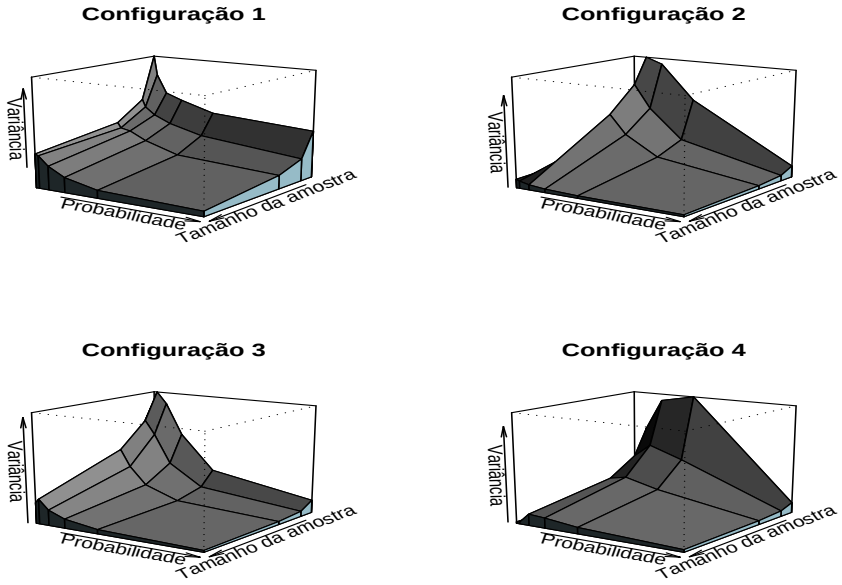
Variância de Q.1, segundo o quantil e o tamanho da amostra, por configuração.

Figura 5.5: *Variância do estimador Q_1.*



Variância de Q.2, segundo o quantil e o tamanho da amostra, por configuração.

Figura 5.6: *Variância do estimador* Q_2.



Variância de Q.3, segundo o quantil e o tamanho da amostra, por configuração.

Figura 5.7: *Variância do estimador* Q_3.

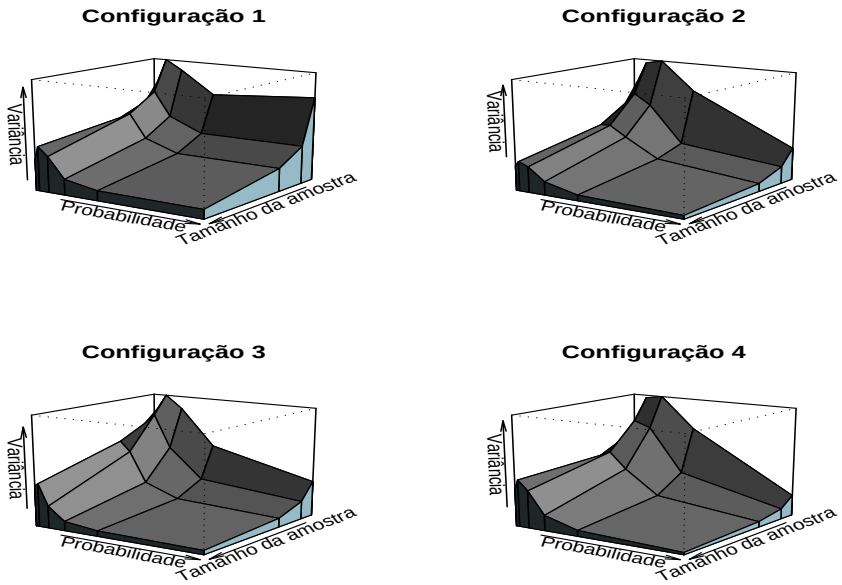


Figura 5.8: *Variância do estimador* Q_4.

Apêndice 2

Histogramas das estimativas dos quantis, segundo o estimador e a configuração.

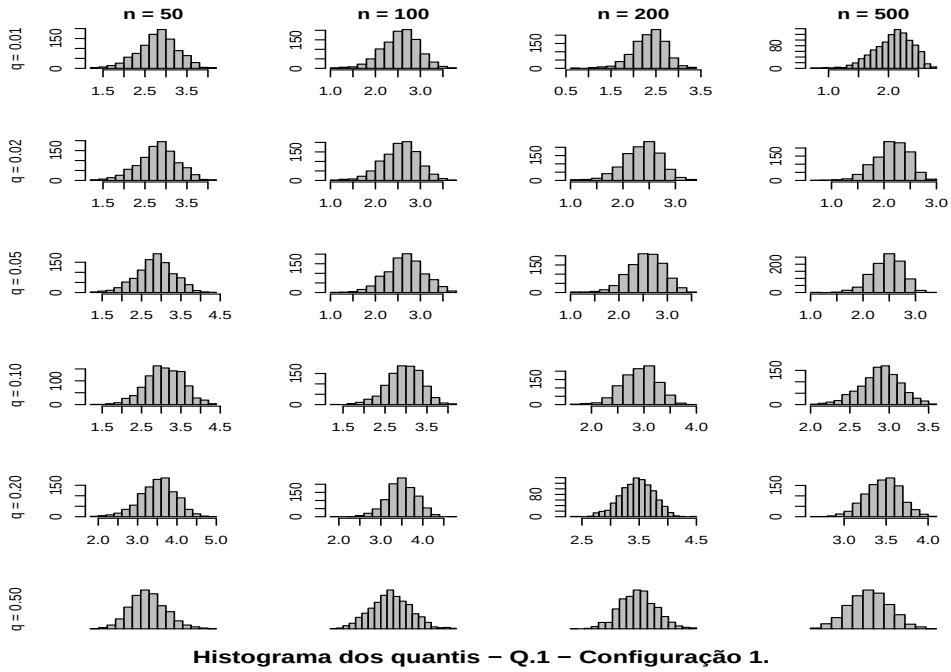


Figura 5.9: *Histograma da estimativa de Q_1, Configuração 1.*

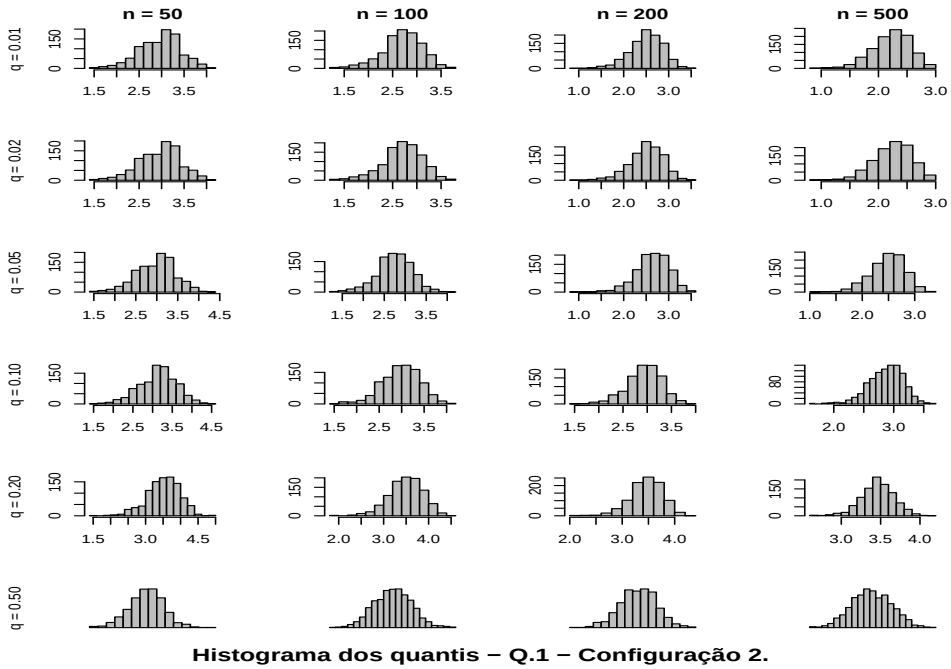


Figura 5.10: *Histograma da estimativa de Q_1, Configuração 2.*

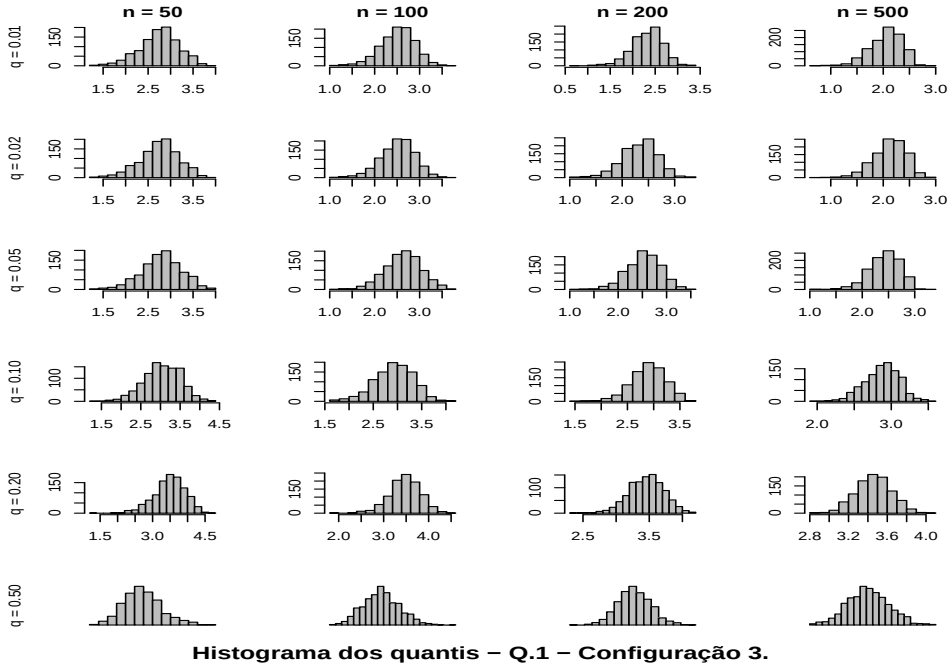


Figura 5.11: *Histograma da estimativa de Q_1, Configuração 3.*

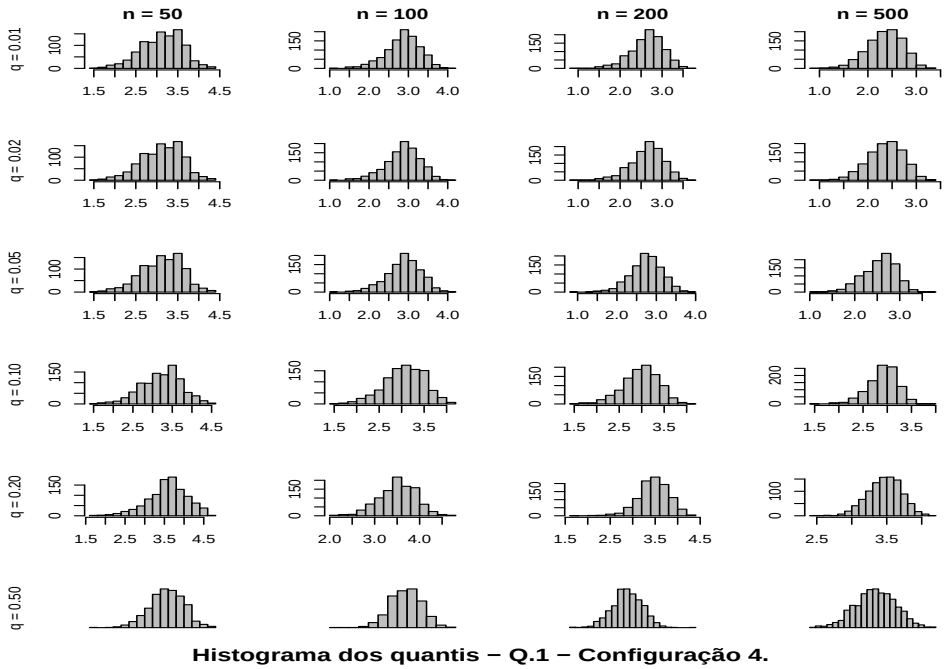


Figura 5.12: *Histograma da estimativa de Q_1, Configuração 4.*

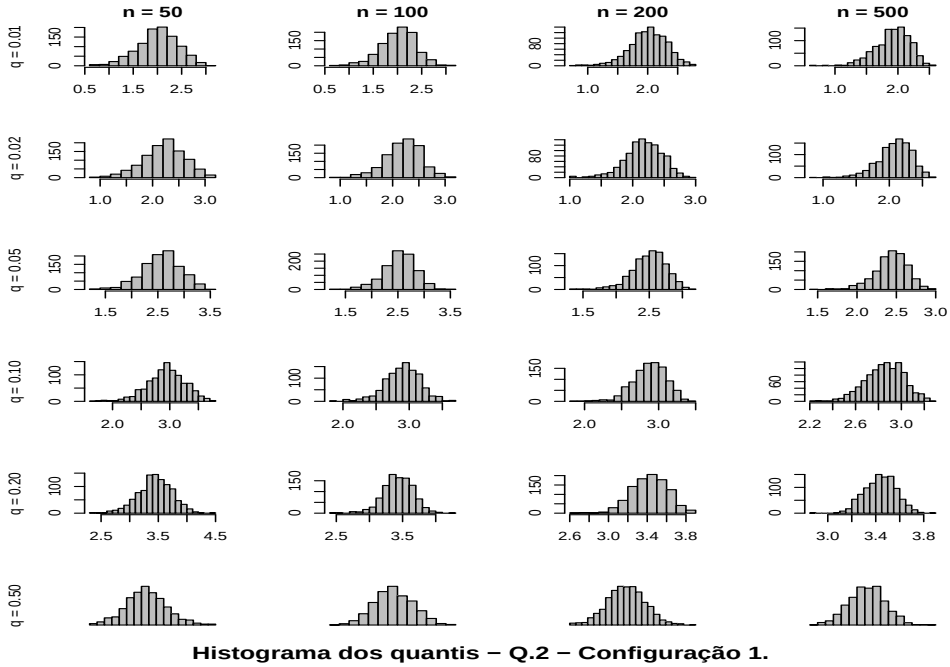


Figura 5.13: *Histograma da estimativa de Q_2, Configuração 1.*

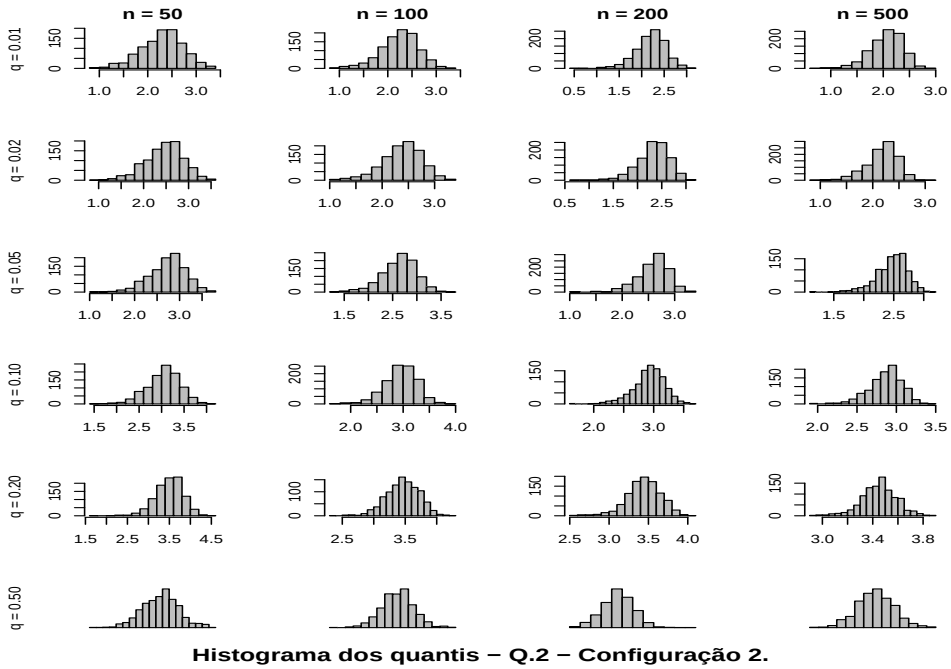


Figura 5.14: *Histograma da estimativa de Q_2, Configuração 2.*

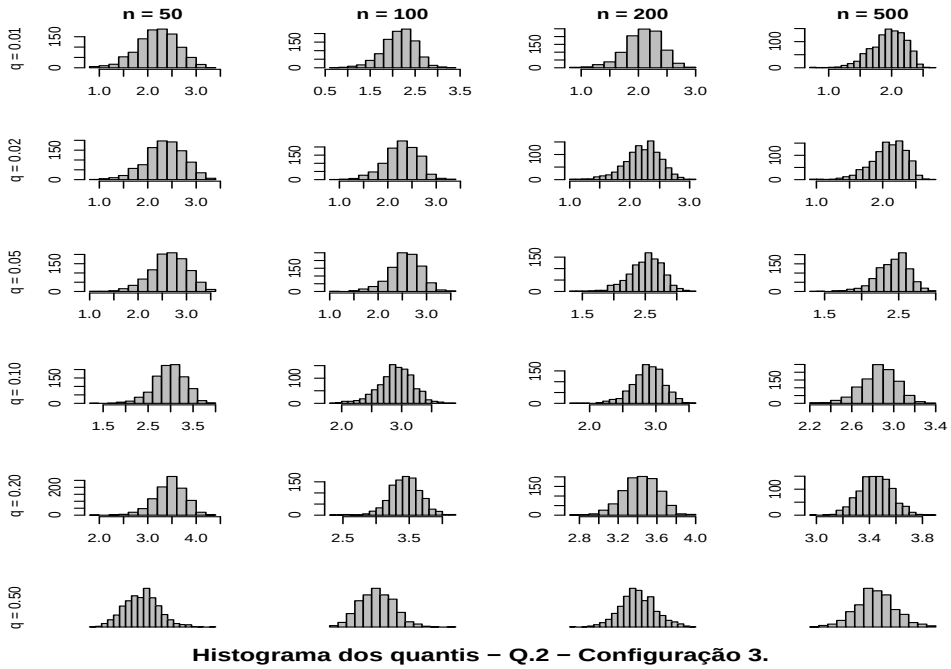


Figura 5.15: *Histograma da estimativa de Q_2, Configuração 3.*

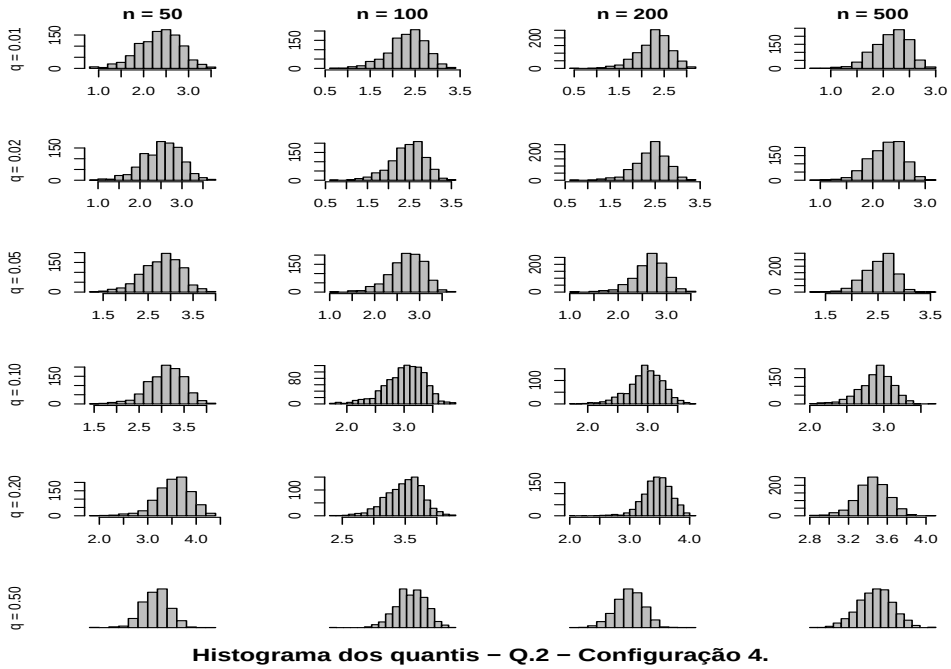


Figura 5.16: *Histograma da estimativa de Q_2, Configuração 4.*

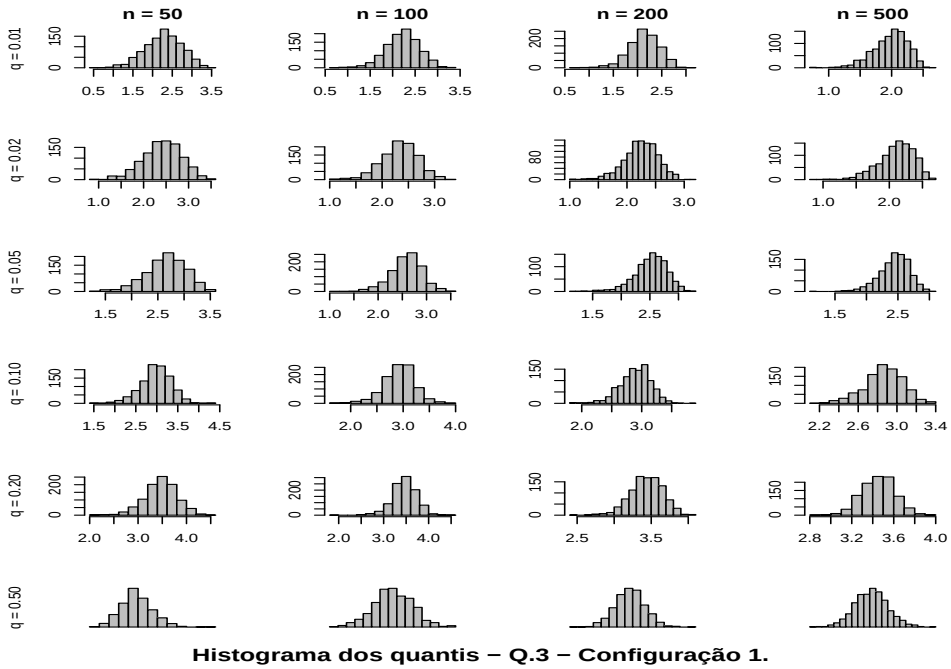


Figura 5.17: *Histograma da estimativa de Q_3, Configuração 1.*

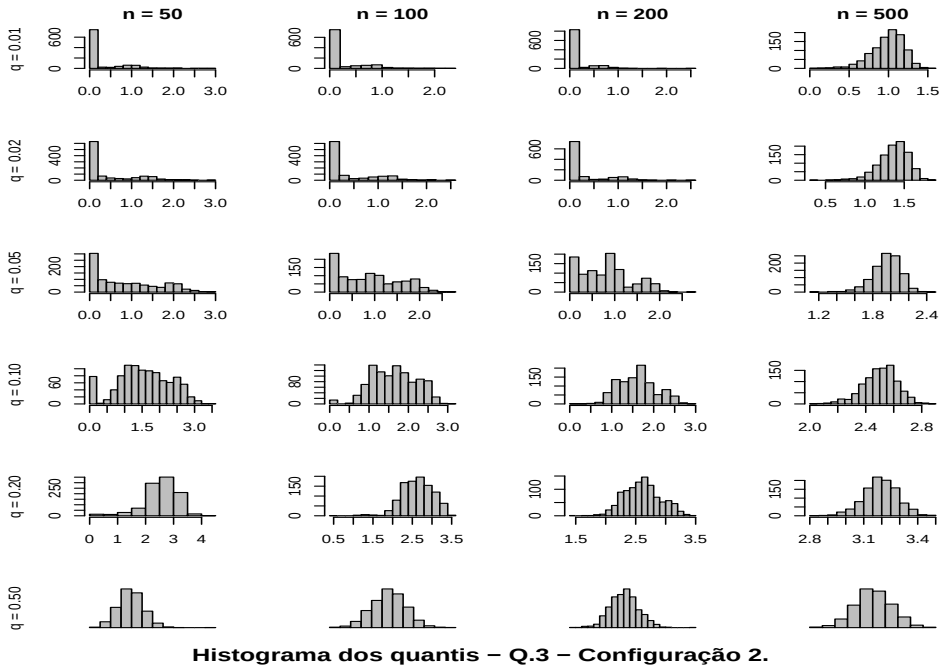


Figura 5.18: *Histograma da estimativa de Q_3, Configuração 2.*

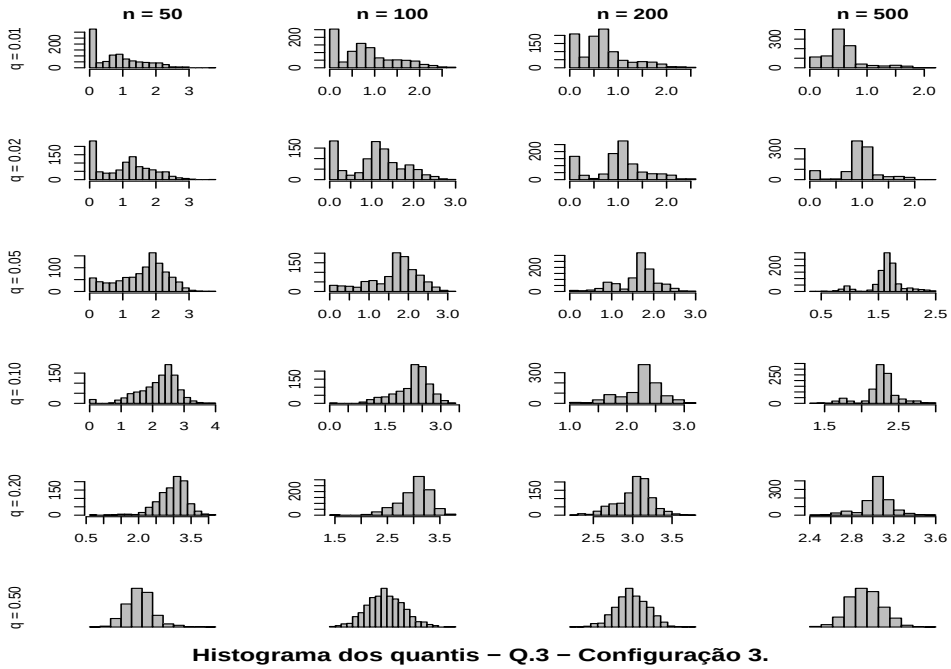


Figura 5.19: *Histograma da estimativa de Q_3, Configuração 3.*

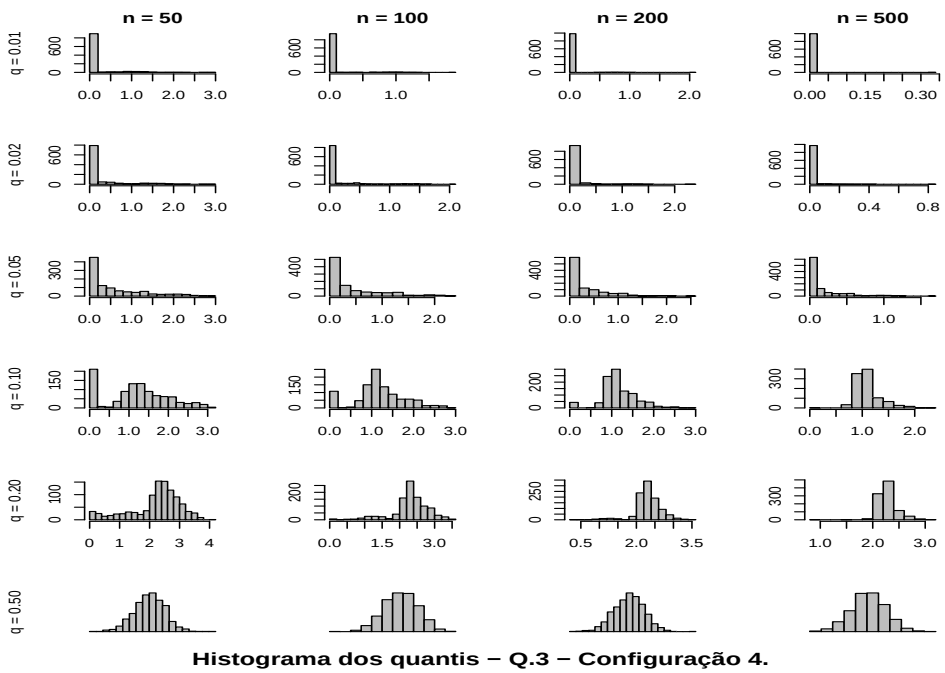


Figura 5.20: *Histograma da estimativa de Q_3, Configuração 4.*

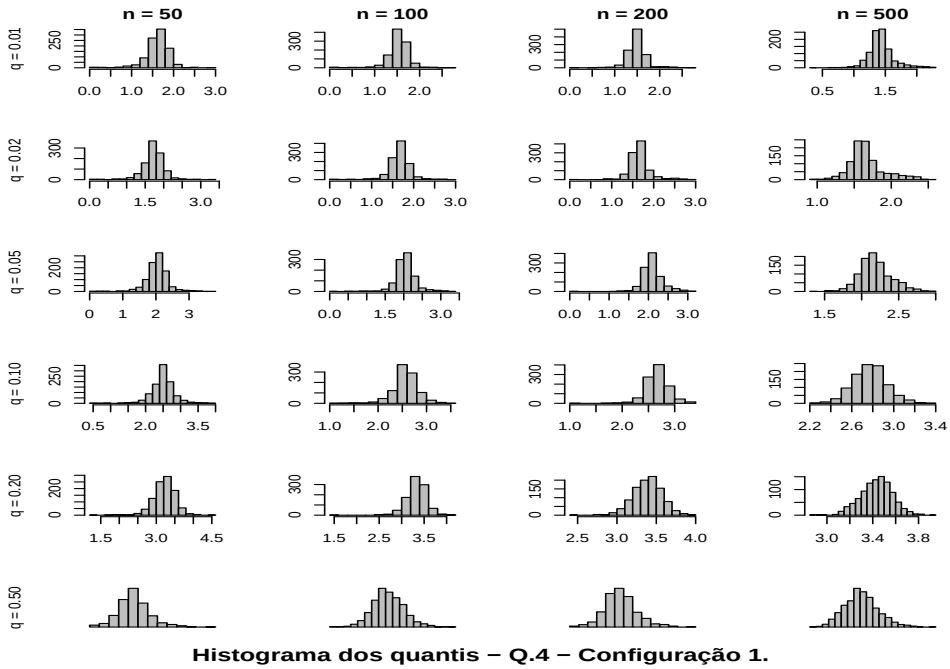


Figura 5.21: *Histograma da estimativa de Q_4, Configuração 1.*

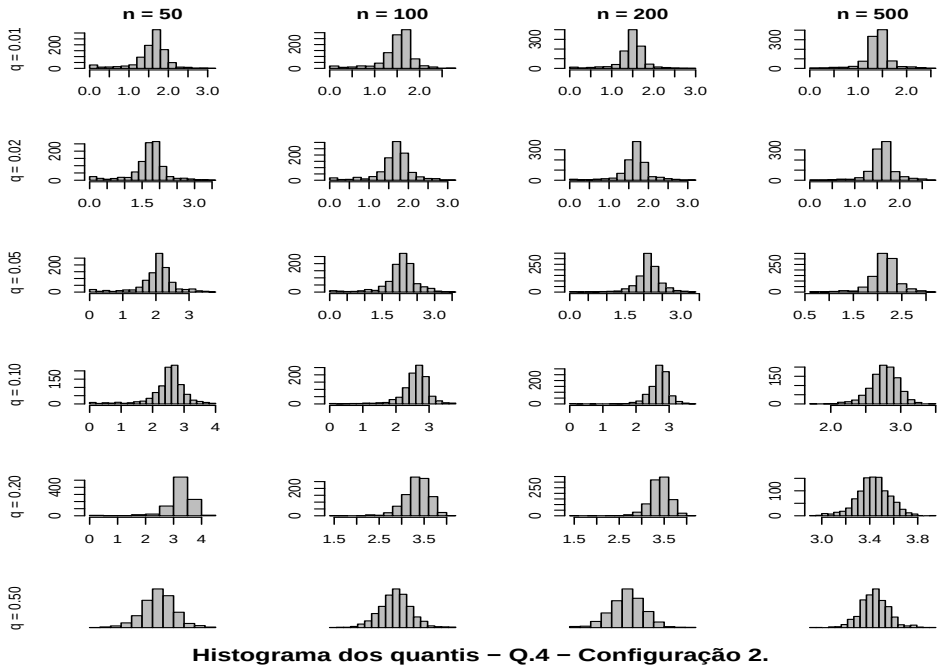


Figura 5.22: *Histograma da estimativa de Q_4, Configuração 2.*

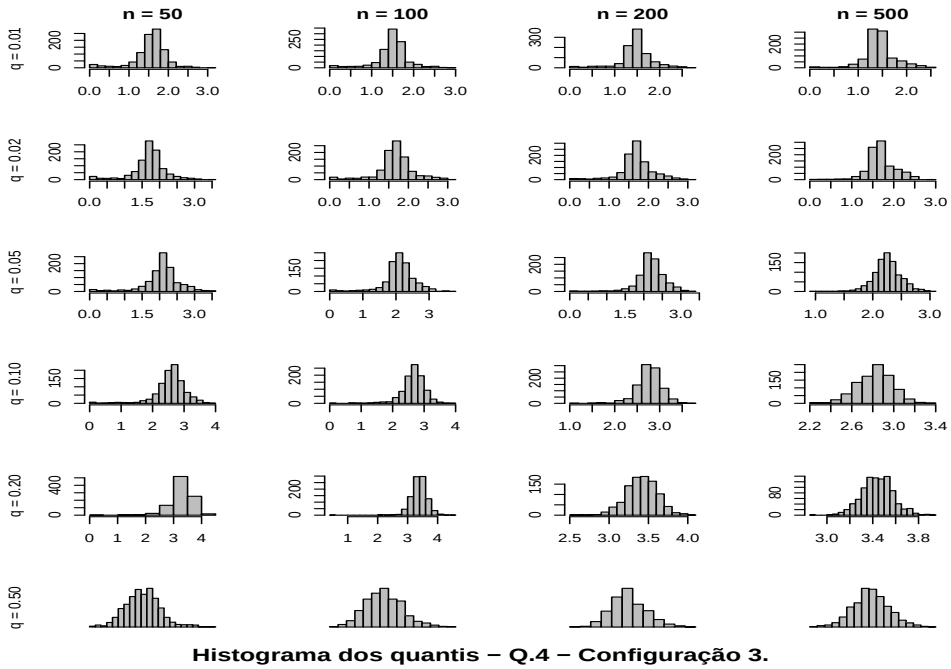


Figura 5.23: *Histograma da estimativa de Q_4, Configuração 3.*

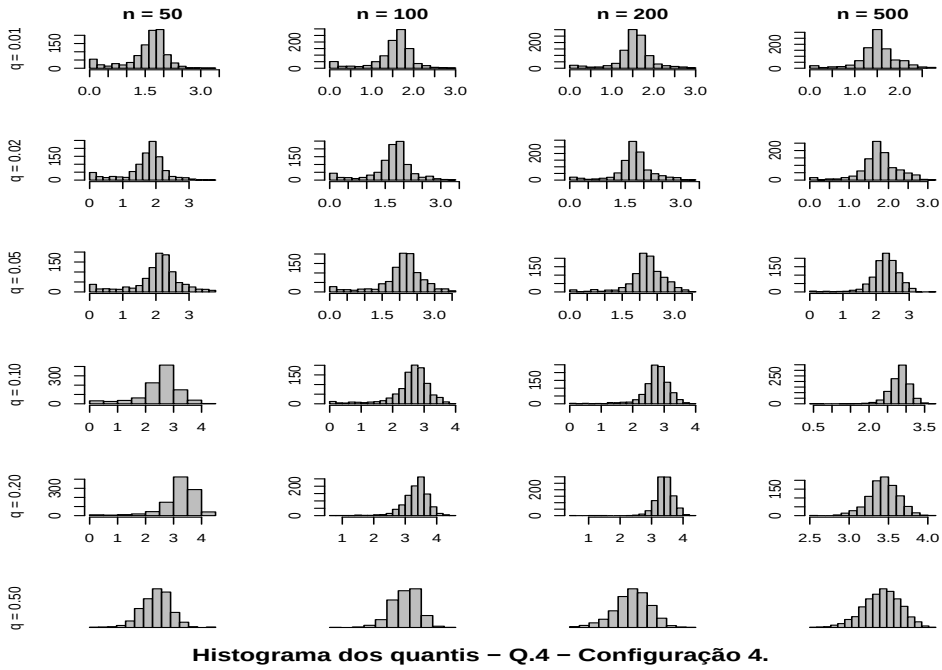


Figura 5.24: *Histograma da estimativa de Q_4, Configuração 4.*