Intervalos de confiança para altos quantis oriundos de distribuições de caudas pesadas

Michel Helcias Montoril

Dissertação apresentada
ao
Instituto de Matemática e Estatística
da
Universidade de São Paulo
para
obtenção do título
de
Mestre em Ciências

Programa: Estatística

Orientador: Profª. Drª. Chang Chiann

Durante o desenvolvimento deste trabalho o autor recebeu auxílio financeiro da CNPq

São Paulo, Fevereiro de 2009

Intervalos de confiança para

altos quantis oriundos de
distribuições de caudas pesadas

Este exemplar corresponde à redação
final da dissertação devidamente corrigida
e defendida por Michel Helcias Montoril
e aprovada pela Comissão Julgadora.

Banca Examinadora:

- Profa. Dra. Chang Chiann (orientador) - IME-USP.
- Profa. Dra. Airlane Pereira Alencar - IME-USP.
- Prof. Dr. Eduardo Fraga Lima de Melo IME-UERJ.

"I can think of younger days when living for my life
Was everything a man could want to do."
Bee Gess

"The ability to simplify means to eliminate
the unnecessary so that the necessary may speak."
Hans Hoffmann

"Man can learn nothing unless he proceeds
from the known to the unknown."
Claude Bernard

"Although this may seem a paradox,
all exact science is dominated by the idea of approximation."
Bertrand Russell

"Le calcul des probabilités n'est au fond
que le bon sens réduit au calcul."
Laplace

"Tudo o que existe no universo é fruto do acaso e da necessidade."
Demócrito

Tale about the cat and the moon

In the beginning there was total darkness.
The silent immensity of the night.
Then she came and everything changed.
It's been a long time since I stopped looking for her.
Now everything is quieter.
I learned that it's better to wait.

She'll come;
When she can.
Or when she wants to.
I know one day she'll come to me.
Otherwise, why would she spend all those hours,
all those nights,
just staring at me?
Nothing else matters.
I'll wait.

But it wasn't always like that.
When I met her, my whole life changed.
I started following her.
I sailed the seas.
Crossed the oceans, for her.
I found myself drifting.
I did everything to find her.
And when I thought I was close.
I was still very far.

I felt lost,
not knowing what to do,
in the middle of all that sea.
The boat was shrinking,
and the world getting smaller and smaller
from all that passion.

Then I changed my life.
I found a steady place and comfortably settled
I thought my proposal was irrefutable.

Once again, she left me.

Desperate, slave of that desire I ran after her.
Jumping from roof to roof.
Prisoner of that attraction,
that was slowly leaving me lonelier.

And time went by.
Now I don't run anymore.
I wait.
Nothing else matters.
I wait.

Pedro Serrazina.

Especialmente à minha mãe, Conceição:

Muito obrigado por tudo que você tem feito por mim, mãe. Você é a principal responsável por tudo de bom que acontece hoje na minha vida. Obrigado pela sua dedicação e preocupação em me manter sempre no caminho do bem. Obrigado por fazer questão que eu sempre tivesse uma boa educação, visando meu futuro e meu bem estar. Obrigado pelas noites em claro, quando eu era pequeno e adoecia. Obrigado por todas as broncas, que com certeza me ajudaram a ser uma pessoa melhor (sei que, se alguma vez você errou, foi tentando acertar). Obrigado por sempre acreditar em mim, como ser humando, como pessoa, como profissional, enfim, como alguém que conseguiria (conseguirá) vencer na vida. Obrigado principalmente por acreditar na minha capacidade, mesmo quando todos duvidaram um dia. Quero que saiba que nunca esqueci, nem esquecerei, tudo o que você fez (e faz até hoje) por mim. Por isso você é a pessoa mais importante da minha vida. Te amo.

♡

Agradecimentos

Gostaria de agradecer a todos que eu acredito que tiveram (direta ou indiretamente) alguma influência neste trabalho, especialmente à:

Deus, por ter me dado a oportunidade de chegar até aqui e por ter me propiciado uma família tão dedicada (pais e irmã).

Aos meus pais, Tiago e Maria Conceição, por toda dedicação, apoio e cuidados na minha vida até os dias de hoje. Por sempre tentarem me dar, aos seus alcances, tudo do bom e do melhor.

À minha irmã, Michelle, por sempre se preocupar comigo, por sempre se dedicar, até mais do que podia, em me ajudar a ser uma boa pessoa. Obrigado por nunca ter esquecido de mim em momento algum. Obrigado por ter me servido como inspiração nos estudos e na determinação. Nunca vi ninguém mais empenhada em vencer na vida do que você.

Às professoras e orientadoras Airlane Pereira Alencar e Chang Chiann, por toda ajuda prestada e dedicação. Diretamente falando, vocês são as principais responsáveis pela conclusão deste trabalho.

Ao Eduardo Fraga de Melo, por ter auxiliado com sugestões valiosas na banca.

Aos professores do IME-USP, pelas ajudas e pelas boas disciplinas ministradas durante o meu programa de mestrado, os quais destaco: Chang Chiann, Mônica Sandoval, Nikolai Kolev, Pedro Morettin, Clélia Toloi, Gilberto Alvarenga e Júlia Pavan.

Aos funcionários do IME-USP, por sempre me atenderem muito bem e por semrpre se empenharem em me ajudar em tudo o que precisei.

A todos que de alguma forma me ajudaram, como Alexandre Galvão, por algumas sugestões dadas, as quais foram úteis para mim.

Aos amigos que me ajudaram bastante na minha chegada a São Paulo, especialmente Rafael Bráz, Núbia, Caio, Iesus, Juvêncio e Moustafá.

Não posso deixar de agradecer ao Juvêncio, pelos memoráveis confrontos de finais de semana no Winning Eleven, onde (acredito eu que propositadamente, apenas para me deixar feliz) ele perdia de mim, na maioria das vezes, por goleadas.

Aos amigos de São Paulo, especialmente ao Rafael Bráz, Tiago (Cara de Tomate), Fabienne Rodrigues, Daniela Caldeirinha, Núbia, Artur, Germán, Estevão, Ivan, João Vinícius, João Celeste, Iesus, Caio, Marcelo, Lane, Ronald, Gilberto, Jacqueline, Juvêncio, Simon, Betsabé, Mariana, Ítalo, Tatiana, Eliana, Alessandra, Luz Marina, Amanda, Catatau, Estéfano, Hommenig, Hugo, Miranda, Rodrigo (Jesus), Hamilton, Nara, Alex (Japa), Thiago Pereira e à toda turma das aulas de forró.

À Elaine Alves, por todo o apoio, carinho e pelos bons momentos que tive o prazer de compartilhar. Espero sempre poder compartilhar, a cada dia, mais e mais coisas legais com você.

A todos os amigos do DEMA-UFC, dentre professores, funcionários e colegas de turma, pelo apoio, insentivo e dedicação que tiveram comigo durante toda a minha graduação. Em especial destaco o amigo João Maurício, por ter me ensinado: boa parte do que aprendi em estatística na graduação, novas palavras bonitas e difíceis, piadas com e sem graça, pensamentos, palavrões e conselhos, que lembro e procuro seguir até hoje.

Aos melhores amigos que fiz nos tempos de graduação: Edson do Carmo, Rafael Bráz, Fabienne, Daniella, Ênio, João Ítalo, Humberto, TT, Joice, Lídia, Suzana, Katiane Marry, Ana Paula, Everton e Everson pelos bons momentos que vocês me propiciaram, pelo apoio e insentivo que foram fundamentais nesta minha caminhada.

Aos meus tios por todo o insentivo e apoio, dentre os quais destaco: Tia Mazé, Tia Fátima, Tio Valderi, Tia Lindaura, Tia Rita e Tia Ester.

Aos meus queridos primos, por todos os bons momentos e pelo carinho que sempre tiveram por mim, especialmente Marquinho, Fabiola, Fabiana, Noelia, Neliane, Neila, Nídia, Ana Paula, Alex, Irisflávia, Osmar Filho e Ronaldo.

Aos bons amigos da juventude em Bela Cruz, pelas risadas, diversões e momentos legais que pude ter, dentre os quais destaco: Marquinho, Eliardo, Denis, Evando, Alexandre Neto, Germano, Rogério, Gugu e Fabinho.

Ao CNPq, pelo apoio financeiro.

E por fim, a todos aqueles que esqueci de mencionar, por de alguma forma terem me ajudado nesta importante fase da minha vida.

Conteúdo

Agradecimentos **v**

Lista de Tabelas **ix**

Lista de Figuras **xiii**

Resumo **xvii**

Abstract **xix**

1 Introdução **1**

1.1 Aplicações diversas da Teoria dos Valores Extremos 1

1.2 Distribuições de caudas pesadas 4

1.3 Objetivos e organização 7

2 Metodologias **9**

2.1 Aproximação pela distribuição normal 10

2.2 Razão de verossimilhanças 13

2.3 Data tilting 15

2.4 Gama generalizada 18

3 Simulações **23**

3.1 Convergência utilizando o método delta 25

3.2 Escolha do limiar 27

3.3 Comparação dos métodos 31

4 Aplicação **45**

5 Conclusões e perspectivas **51**

A Demonstrações **53**

A.1 Cálculo de $\hat{\gamma}_n$ e $\hat{c}_n$ 53

A.2 Relação entre as funções de distribuição e quantil 53

A.3 Multiplicadores de Lagrange para obter γ e c no método da razão de verossimilhanças 54

A.4 Cálculo dos estimadores de γ e c no método *data tilting*. 56

A.5 Multiplicadores de Lagrange para o vetor de pesos do método *data tilting* 57

A.6 Relação entre as distribuições Gama e Gama Generalizada 59

B Método para a seleção do valor de k em $X_{n-k,n}$ **61**

B.1 Pelo erro quadrático médio assintótico do estimador de Hill 61

Referências Bibliográficas **65**

Lista de Tabelas

2.1 Distribuições que podem ser escritas como casos particulares da Gama Generalizada apresentadas em Stacy e Mihram (1965) 19

3.1 Número de casos onde a estimativa da matriz de informação não foi positiva definida, de acordo com o tamanho amostral, em um total de 10000 amostras. 26

3.2 Estimativas médias dos parâmetros da distribuição Gama Generalizada, para cada tamanho amostral, com $\beta = 0,3$, $\zeta = 1$ e $\alpha = 1$. 26

3.3 Estimativas médias para os quantis de ordem $0,99$ e $0,999$, para 10000 amostras geradas da distribuição Weibull, variando o tamanho amostral, segundo o método delta (gama generalizada), com $q_{0,99} = 162,4871$ e $q_{0,999} = 627,7545$. 28

3.4 Probabilidades de cobertura e amplitudes médias dos intervalos de 90% de confiança para os quantis de ordem $0,99$ e $0,999$, para 10000 amostras geradas da distribuição Weibull, variando o tamanho amostral, segundo o método delta (gama generalizada), com $q_{0,99} = 162,4871$ e $q_{0,999} = 627,7545$. 29

3.5 Limites inferior e superior médios dos intervalos de 90% de confiança para os quantis de ordem $0,99$ e $0,999$, para 10000 amostras geradas da distribuição Weibull, variando o tamanho amostral, segundo o método delta (gama generalizada), com $q_{0,99} = 162,4871$ e $q_{0,999} = 627,7545$. 29

3.6 Estatísticas descritivas (mínimo, primeiro quartil, mediana, média, terceiro quartil, máximo e desvio padrão) dos valores de k que minimizam o erro quadrático médio assintótico do estimador de Hill, k_2, para 10000 amostras de tamanhos 1000 e 2000 das distribuições Weibull e Fréchet. 30

3.7 Probabilidades de coberturas e amplitudes médias dos intervalos com 90% de confiança para o quantil de ordem $0,99$ ($x_{0,01} = 162,4871$), para 10000 amostras de tamanhos 1000 e 2000 da distribuição Weibull, utilizando k_1 e k_2. 38

3.8 Probabilidades de coberturas e amplitudes médias dos intervalos com 90% de confiança para o quantil de ordem $0,999$ ($x_{0,001} = 627,7545$), para 10000 amostras de tamanhos 1000 e 2000 da distribuição Weibull, utilizando k_1 e k_2. 38

3.9 Limites inferior e superior médios dos intervalos com 90% de confiança para o quantil de ordem $0,99$ ($x_{0,01} = 162,4871$), para 10000 amostras de tamanhos 1000 e 2000 da distribuição Weibull, utilizando k_1 e k_2. 38

3.10 Limites inferior e superior médios dos intervalos com 90% de confiança para o quantil de ordem $0,999$ ($x_{0,001} = 627,7545$), para 10000 amostras de tamanhos 1000 e 2000 da distribuição Weibull, utilizando k_1 e k_2. 39

3.11 Probabilidades de coberturas e amplitudes médias dos intervalos com 90% de confiança para o quantil de ordem $0,99$ ($x_{0,01} = 99,49916$), para 10000 amostras de tamanhos 1000 e 2000 da distribuição Fréchet, utilizando k_1 e k_2. 39

3.12 Probabilidades de coberturas e amplitudes médias dos intervalos com 90% de confiança para o quantil de ordem $0,999$ ($x_{0,001} = 999,49992$), para amostras de tamanhos 1000 e 2000 da distribuição Fréchet, utilizando k_1 e k_2. 39

3.13 Limites inferior e superior médios dos intervalos com 90% de confiança para o quantil de ordem $0,99$ ($x_{0,01} = 99,49916$), para amostras de tamanhos 1000 e 2000 da distribuição Fréchet, utilizando k_1 e k_2. 40

3.14 Limites inferior e superior médios dos intervalos com 90% de confiança para o quantil de ordem $0,999$ ($x_{0,001} = 999,49992$), para amostras de tamanhos 1000 e 2000 da distribuição Fréchet, utilizando k_1 e k_2. 40

3.15 Probabilidades de coberturas e amplitudes médias dos intervalos com 90% de confiança para o quantil de ordem $0,99$ ($x_{0,01} = 162,4871$), para 10000 amostras de tamanhos 1000 e 2000 da distribuição Weibull, utilizando k_2 nos métodos da aproximação pela normal, razão de verossimilhanças e *data tilting*. 41

3.16 Probabilidades de coberturas e amplitudes médias dos intervalos com 90% de confiança para o quantil de ordem $0,999$ ($x_{0,001} = 627,7545$), para 10000 amostras de tamanhos 1000 e 2000 da distribuição Weibull, utilizando k_2 nos métodos da aproximação pela normal, razão de verossimilhanças e *data tilting*. 41

3.17 Probabilidades de coberturas e amplitudes médias dos intervalos com 90% de confiança para o quantil de ordem $0,99$ ($x_{0,01} = 99,49916$), para 10000 amostras de tamanhos 1000 e 2000 da distribuição Fréchet, utilizando k_1 nos métodos da aproximação pela normal, razão de verossimilhanças e *data tilting*. 41
3.18 Probabilidades de coberturas e amplitudes médias dos intervalos com 90% de confiança para o quantil de ordem $0,999$ ($x_{0,001} = 999,49992$), para 10000 amostras de tamanhos 1000 e 2000 da distribuição Fréchet, utilizando k_1 nos métodos da aproximação pela normal, razão de verossimilhanças e *data tilting*. 41
3.19 Limites inferior e superior médios dos intervalos com 90% de confiança para o quantil de ordem $0,99$ ($x_{0,01} = 162,4871$), para 10000 amostras de tamanhos 1000 e 2000 da distribuição Weibull, utilizando k_2 nos métodos da aproximação pela normal, razão de verossimilhanças e *data tilting*. 42
3.20 Limites inferior e superior médios dos intervalos com 90% de confiança para o quantil de ordem $0,999$ ($x_{0,001} = 627,7545$), para 10000 amostras de tamanhos 1000 e 2000 da distribuição Weibull, utilizando k_2 nos métodos da aproximação pela normal, razão de verossimilhanças e *data tilting*. 42
3.21 Limites inferior e superior médios dos intervalos com 90% de confiança para o quantil de ordem $0,99$ ($x_{0,01} = 99,49916$), para 10000 amostras de tamanhos 1000 e 2000 da distribuição Fréchet, utilizando k_1 nos métodos da aproximação pela normal, razão de verossimilhanças e *data tilting*. 42
3.22 Limites inferior e superior médios dos intervalos com 90% de confiança para o quantil de ordem $0,999$ ($x_{0,001} = 999,49992$), para 10000 amostras de tamanhos 1000 e 2000 da distribuição Fréchet, utilizando k_1 nos métodos da aproximação pela normal, razão de verossimilhanças e *data tilting*. 42

4.1 Estatísticas descritivas referentes ao pagamento de indenizações de seguros de incêndio no Brasil em 2003 45
4.2 Intervalos de 90% de confiança para os quantis de ordem $0,99$ e $0,999$ das indenizações dos seguros de incêndios. 50

Lista de Figuras

3.1 Função densidade de probabilidade da distribuição Weibull com parâmetros $\beta = 0,3$ e $\alpha = 1$ 24

3.2 Função densidade de probabilidade da distribuição Fréchet com parâmetro $\alpha = 1$ . . 24

3.3 Histogramas das 10000 estimativas de β comparadas com a distribuição normal . . . 26

3.4 Histogramas das 10000 estimativas de ζ comparadas com a distribuição normal . . . 27

3.5 Histogramas das 10000 estimativas de α comparadas com a distribuição normal . . . 27

3.6 Histogramas das 10000 estimativas do quantil de ordem $0,99$ comparadas com a distribuição normal 28

3.7 Histogramas das 10000 estimativas do quantil de ordem $0,999$ comparadas com a distribuição normal 28

3.8 Probabilidades de cobertura dos intervalos com 90% de confiança para os quantis de ordem $0,99$ ($x_{0,01}$) – lado esquerdo – e $0,999$ ($x_{0,001}$) – lado direito – da distribuição Weibull, comparando as escolhas k_1 e k_2, para os métodos da aproximação pela normal, razão de verossimilhanças e *data tilting*. 32

3.9 Amplitudes médias dos intervalos com 90% de confiança para os quantis de ordem $0,99$ ($x_{0,01}$) – lado esquerdo – e $0,999$ ($x_{0,001}$) – lado direito – da distribuição Weibull, comparando as escolhas k_1 e k_2, para os métodos da aproximação pela normal, razão de verossimilhanças e *data tilting*. 33

3.10 Limites inferiores e superiores médios dos intervalos com 90% de confiança para os quantis de ordem $0,99$ ($x_{0,01}$) – lado esquerdo – e $0,999$ ($x_{0,001}$) – lado direito – da distribuição Weibull, comparando as escolhas k_1 e k_2, para os métodos da aproximação pela normal, razão de verossimilhanças e *data tilting*. 34

3.11 Probabilidades de cobertura dos intervalos com 90% de confiança para os quantis de ordem $0,99$ ($x_{0,01}$) – lado esquerdo – e $0,999$ ($x_{0,001}$) – lado direito – da distribuição Fréchet, comparando as escolhas k_1 e k_2, para os métodos da aproximação pela normal, razão de verossimilhanças e *data tilting*. 35

3.12 Amplitudes médias dos intervalos com 90% de confiança para os quantis de ordem $0,99$ ($x_{0,01}$) – lado esquerdo – e $0,999$ ($x_{0,001}$) – lado direito – da distribuição Fréchet, comparando as escolhas k_1 e k_2, para os métodos da aproximação pela normal, razão de verossimilhanças e *data tilting*. 36

3.13 Limites inferiores e superiores médios dos intervalos com 90% de confiança para os quantis de ordem $0,99$ ($x_{0,01}$) – lado esquerdo – e $0,999$ ($x_{0,001}$) – lado direito – da distribuição Fréchet, comparando as escolhas k_1 e k_2, para os métodos da aproximação pela normal, razão de verossimilhanças e *data tilting*. 37

3.14 Probabilidades de cobertura dos intervalos com 90% de confiança para os quantis de ordem $0,99$ ($x_{0,01}$) e $0,999$ ($x_{0,001}$) da distribuição Weibull, comparando os métodos da aproximação pela normal, razão de verossimilhanças, *data tilting* e gama generalizada. 43

3.15 Probabilidades de cobertura dos intervalos com 90% de confiança para os quantis de ordem $0,99$ ($x_{0,01}$) e $0,999$ ($x_{0,001}$) da distribuição Fréchet, comparando os métodos da aproximação pela normal, razão de verossimilhanças e *data tilting*. 44

4.1 Dados referentes ao pagamento de indenizações de seguros de incêndio no Brasil em 2003. 46

4.2 Histograma dos dados referentes ao pagamento de indenizações de seguros de incêndio no Brasil em 2003. 46

4.3 Comparação entre gráficos log–log para distribuições com cauda de variação regular e distribuições subexponenciais que não possuem cauda de variação regular. 47

4.4 Gráfico log-log aplicado aos dados referentes ao pagamentos de indenizações de seguros de incêndio no Brasil em 2003. 49

4.5 Gráfico das estimativas de Hill aplicado aos dados referentes ao pagamentos de indenizações de seguros de incêndio no Brasil em 2003. 49

4.6 Gráfico da razão entre máximos e somas aplicado aos dados referentes ao pagamentos de indenizações de seguros de incêndio no Brasil em 2003. 49

4.7 (a) Gráfico da distribuição empírica com a função de distribuição ajustada pela gama generalizada e (b) QQ-plot dos dados comparados à gama generalizada 50

Resumo

Este trabalho tem como objetivo calcular intervalos de confiança para altos quantis oriundos de distribuições de caudas pesadas. Para isso, utilizamos os métodos da aproximação pela distribuição normal, razão de verossimilhanças, *data tilting* e gama generalizada. Obtivemos, através de simulações, que os intervalos calculados a partir do método da gama generalizada apresentam probabilidades de cobertura bem próximas do nível de confiança, com amplitudes médias menores do que os outros três métodos, para dados gerados da distribuição Weibull. Todavia, para dados gerados da distribuição Fréchet, o método da razão de verossimilhanças fornece os melhores intervalos. Aplicamos os métodos utilizados neste trabalho a um conjunto de dados reais, referentes aos pagamentos de indenizações, em reais, de seguros de incêndio, de um determinado grupo de seguradoras no Brasil, no ano de 2003.

Palavras-chave: eventos extremos, distribuições de caudas pesadas, altos quantis, intervalos de confiança.

Abstract

In this work, confidence intervals for high quantiles from heavy-tailed distributions were computed. More specifically, four methods, namely, normal approximation method, likelihood ratio method, data tilting method and generalised gamma method are used. A simulation study with data generated from Weibull distribution has shown that the generalised gamma method has better coverage probabilities with the smallest average length intervals. However, from data generated from Fréchet distribution, the likelihood ratio method gives the better intervals. Moreover, the methods used in this work are applied on a real data set from 1758 Brazilian fire claims.

Palavras-chave: extremal events, heavy-tailed distributions, high quantiles, confidence intervals.

Capítulo 1

Introdução

Nos últimos anos, vem crescendo bastante o interesse na previsão de eventos extremos nas mais diversas áreas do conhecimento. Muitas vezes tais eventos podem ser caracterizados por valores muito altos (ou muito baixos) de alguma variável aleatória quantitativa, os quais são denominados valores extremos. O estudo a respeito de tais valores pertence a uma importante área da estatística, conhecida por teoria dos valores extremos (TVE). A análise de valores extremos comumente se dá por meio do estudo da função de distribuição $F(x) := \mathbb{P}(X \leq x)$, da função de sobrevivência (também conhecida por probabilidade da cauda) $\bar{F}(x) = 1 - F(x)$ ou do quantil $x_p = \inf\{x : F(x) \geq p\}$, de alguma variável X de interesse. Neste trabalho estudaremos o comportamento do quantil, quando p está próximo de 1, ou seja, estaremos interessados na estimação de altos quantis.

Particularmente, estimativas pontuais podem não fornecer resultados muito próximos do verdadeiro valor do quantil de interesse, daí surge a preocupação em se obter intervalos de confiança, os quais fornecerão um conjunto de valores plausíveis, com base na amostra, a um determinado nível de confiança.

1.1 Aplicações diversas da Teoria dos Valores Extremos

Nesta seção comentaremos brevemente como as técnicas da teoria dos valores extremos vêm sendo utilizadas em diversas áreas, exibindo três exemplos diferentes, os quais também são comentados em Beirlant *et al.* (2004), com o intuito de mostrar a diversidade de aplicações.

Em hidrologia é comum o estudo de determinadas grandezas hidrológicas (como precipitações, vazões, evaporação etc), as quais, quando observadas ao longo do tempo, eventualmente, apresentam variações periódicas, geralmente anuais. Uma forma comumente utilizada na área para eliminar a dependência dos dados (como a periodicidade) e estudar valores extremos é utilizar os máximos anuais. A partir de então, pode-se pensar em avaliar o tempo de ocorrência de uma determinada

grandeza hidrológica a partir da estimação do que se chama de período de retorno de T anos. Pode-se ainda ter o interesse em estudar a magnitude dessa grandeza para um dado período de tempo T, a qual é denominada por nível de retorno de T anos. Se pensarmos no máximo anual de uma determinada grandeza hidrológica como sendo uma variável X, teremos que o período de retorno corresponde a

$$T(x) = \frac{1}{\mathbb{P}(X > x)},$$

ou seja, o inverso da probabilidade de sobrevivência de uma determinada grandeza X aplicada em uma dada magnitude x. O nível de retorno de T anos, x_T, corresponde ao $(1 - 1/T)$-ésimo quantil de X, com

$$x_T = \inf\left\{x : F(x) \geq 1 - \frac{1}{T}\right\}$$

em que $F(x) = \mathbb{P}(X \leq x)$. Em termos gerais dizemos que x_T corresponde ao nível que será excedido em média, a cada T anos, ou ainda, o nível que será excedido com probabilidade $1/T$ em um ano qualquer. É possível observar que o período de retorno e o nível de retorno são diretamente relacionados. Comumente tem-se o interesse em obter o nível de retorno de $T = 100$ anos, x_{100} (ou seja, o $0,99$-quantil, pois $1 - 1/100 = 0,99$, dos máximos anuais), embora as estimações frequentemente sejam baseadas em curtos períodos de tempo, como ocorre no conjunto de dados aplicados em Van Noortwijk (1999). Beirlant *et al.* (2004) citam exemplos que mostram conseqüências desastrosas quando o nível de retorno é excedido, o que motiva o interesse na estimação intervalar (intervalos de confiança) em vez da pontual. Existem alguns trabalhos que focam no uso de intervalos de confiança para níveis de retorno, como, por exemplo, Rust *et al.* (2006), onde métodos bootstrap são utilizados para calcular intervalos de confiança para níveis de retorno de enchentes.

Outra área em que a teoria dos valores extremos é bastante utilizada é em finanças. Uma quantidade considerável de séries temporais financeiras consiste de preços de ativos financeiros. Um dos objetivos em finanças é a avaliação de riscos de uma carteira de ativos financeiros, os quais são, geralmente, medidos por meio da variação dos preços de tais ativos. Supondo que essa série temporal seja igualmente espaçada, ou seja, o intervalo de tempo entre as observações não varia (as observações são obtidas, por exemplo, diariamente), podemos denotar o preço de um determinado ativo no instante t por P_t. Os principais tipos de variação dos preços são denominados de retorno líquido simples, retorno bruto simples e log-retorno. O retorno líquido simples no instante

t é definido como a variação relativa entre os preços consecutivos

$$R_t = \frac{P_t - P_{t-1}}{P_{t-1}} = \frac{P_t}{P_{t-1}} - 1$$

e o retorno bruto simples corresponde a $1 + R_t$. O log-retorno (que costuma ser o mais utilizado) é escrito como

$$r_t = \log(1 + R_t).$$

Como já mencionamos, freqüentemente se está interessado em avaliar riscos financeiros de mercado, uma vez que os mesmos estão ligados às perdas potenciais relacionadas ao comportamento do mercado. Uma das medidas mais utilizadas para se avaliar tal risco é conhecida por `VaR` (valor em risco). De acordo com Tsay (2002), do ponto de vista de uma instituição financeira, o `VaR` pode ser definido como a perda máxima de uma posição financeira durante um dado período de tempo com uma certa probabilidade, tratando-se como uma medida de perda associada a um evento raro sob condições normais de mercado. Podemos pensar ainda, no `VaR` como sendo o p-ésimo quantil do retorno (em geral, um alto quantil dessa distribuição é de interesse do mercado). Existem algumas formas de se estimar o `VaR`, dentre elas utilizando-se a teoria dos valores extremos. Não entraremos em detalhes sobre as técnicas utilizadas para o cálculo do valor em risco. Maiores detalhes sobre `VaR` podem ser obtidos em Tsay (2002), onde há também mais informações sobre outros tipos de retornos. Uma forma de se ter um maior controle sob as estimativas do `VaR` seria com a utilização de intervalos de confiança. Chan *et al.* (2007) obtêm intervalos de confiança para o valor em risco, baseados em modelos GARCH com inovações de caudas pesadas.

Uma das mais importantes aplicações de valor extremo pode ser encontrada em seguros. Por exemplo, em seguros de vida, de incêndios e de automóveis, ocorrem muitos sinistros. Segundo Beirlant *et al.* (2004), incêndios industriais, especificamente, causam vários efeitos colaterais na perda de propriedade, desempregos temporários e perdas de contratos. Alguns, ocasionalmente, incluem grandes sinistros, o que coloca em risco a solvência de um portfolio ou, até mesmo, de uma parte substancial da companhia. Daí a necessidade de previsão, por parte das companhias, dos altos quantis da distribuição do valor dos sinistros, o que motiva o desenvolvimento de técnicas apropriadas para a estimação dos mesmos. Existe na literatura, uma quantidade considerável de trabalhos que visam a estimação (tanto pontual quanto intervalar) de altos quantis. Peng e Qi (2006) calculam intervalos de confiança para altos quantis de distribuições de caudas pesadas, usando três métodos diferentes, os quais serão utilizados e discutidos neste trabalho. Além disso, as companhias

de seguros, visando diminuir suas responsabilidades na aceitação de um risco considerado excessivo ou perigoso, cedem às resseguradoras uma parte da responsabilidade e do prêmio recebido. Em um contrato de resseguro de excesso de danos, a resseguradora paga pelo montante excedente, a um determinado limite dos sinistros. A distribuição da cauda superior do valor dos sinistros é de maior interesse para determinar o prêmio líquido de um contrato de resseguro. Várias novas direções na teoria dos valores extremos foi influenciada pelos métodos desenvolvidos na literatura atuarial.

Outros exemplos de aplicações da teoria dos valores extremos podem ser encontrados em Beirlant *et al.* (2004). Coles (2001) também cita alguns trabalhos de outras áreas que fazem uso da teoria dos valores extremos.

1.2 Distribuições de caudas pesadas

Uma vez que nosso trabalho tem como objetivo principal a estimação intervalar de altos quantis oriundos de distribuições de caudas pesadas, apresentaremos algumas definições e resultados sobre tais distribuições.

Não existe uma definição precisa com relação a distribuições de caudas pesadas. Contudo, uma das definições mais utilizadas é baseada no coeficiente de curtose da variável aleatória. Neste caso, diz-se que uma variável aleatória X possui distribuição de caudas pesadas se a mesma possuir o coeficiente de curtose superior ao da distribuição normal, ou seja, para μ_X e σ_X sendo a média e o desvio-padrão de X, respectivamente, tem-se que

$$\mathbb{E}\left[\left(\frac{X-\mu_X}{\sigma_X}\right)^4\right] > 3.$$

Todavia, com esta definição não temos como afirmar se uma variável que não possui o quarto momento tem distribuição de caudas pesadas. Desta forma, utilizaremos a definição de Sigman (1999), ou seja, dada uma variável X, com função de distribuição $F(x) = \mathbb{P}(X \leq x)$ e função de sobrevivência (ou probabilidade da cauda) $\bar{F}(x) = 1 - F(x)$, diremos que X possui cauda pesada se $\bar{F}(x) \geq 0 \quad \forall x \geq 0$ e

$$\lim_{x\to\infty} \mathbb{P}(X > x+y | X > x) = \lim_{x\to\infty} \frac{\bar{F}(x+y)}{\bar{F}(x)} = 1, \quad y \geq 0. \tag{1.1}$$

Desse modo, com base em (1.1), é possível verificarmos que toda distribuição de cauda pesada possui função geradora de momentos infinita, ou seja,

$$\mathbb{E}\left(e^{tX}\right) = \infty.$$

Denominaremos a classe de distribuições de cauda pesada por $\mathcal{L}$ e utilizaremos a notação $X \in \mathcal{L}$. Dentro dessa classe de distribuições existem várias outras subclasses, dentre elas as distribuições subexponeciais $\mathcal{S}$ e as distribuições com caudas de variação regular $\mathcal{R}$, as quais serão utilizadas no decorrer do nosso trabalho. Vamos então às definições das duas subclasses.

Definição 1 *Uma variável aleatória X, positiva, é dita ter distribuição subexponencial (com notação $X \in \mathcal{S}$), se para todo $n \geq 2$,*

$$\lim_{x\to\infty} \frac{\bar{F}^{n*}(x)}{\bar{F}(x)} = n, \tag{1.2}$$

onde $\bar{F}^{n*}(x) = \mathbb{P}(X_1 + X_2 + \ldots + X_n > x)$.

O nome subexponencial se dá pelo fato desse tipo de distribuição possuir caudas com decaimento mais lento do que a cauda de qualquer distribuição exponencial com média $1/\lambda > 0$, isto é,

$$\lim_{x\to\infty} \frac{\bar{F}(x)}{e^{-\lambda x}} = \lim_{x\to\infty} e^{\lambda x}\bar{F}(x) = \infty.$$

A distribuição Weibull, que será usada no decorrer do texto, possui cauda pesada e pertence à classe de distrbiuição subexponencial quando seu parâmetro de forma é inferior a 1. A função de sobrevivência da Weibull é definida por $\bar{F}(x) = e^{-(\frac{x}{\alpha})^{\beta}}$, $x > 0$, onde α ($\alpha > 0$) é conhecido como parâmetro de escala e β ($\beta > 0$) como de parâmetro de forma. Suponhamos que $0 < \beta < 1$. Assim,

$$\begin{aligned}
\lim_{x\to\infty} \frac{\bar{F}(x)}{e^{-\lambda x}} &= \lim_{x\to\infty} e^{\lambda x} e^{-(\frac{x}{\alpha})^{\beta}} = \lim_{x\to\infty} e^{\lambda x - (\frac{1}{\alpha})^{\beta} x^{\beta}} \\
&= \exp\left\{ \lim_{x\to\infty} \left[x \left(\lambda - \frac{1}{\alpha^{\beta}} x^{\beta-1} \right) \right] \right\} \\
&= \exp\left\{ \left[\lim_{x\to\infty} x \right] \left[\lim_{x\to\infty} \left(\lambda - \frac{1}{\alpha^{\beta}} x^{\beta-1} \right) \right] \right\}
\end{aligned}$$

Observe que, como estamos supondo que $0 < \beta < 1$, $\lim_{x\to\infty}(\lambda - \frac{1}{\alpha^{\beta}} x^{\beta-1}) = \lambda$. Logo, teremos

$$\lim_{x\to\infty} e^{\lambda x} e^{-(\frac{x}{\alpha})^{\beta}} = \exp\left\{ \lim_{x\to\infty} x\lambda \right\} = \infty.$$

Antes de definirmos distribuições com cauda de variação regular, falaremos brevemente sobre

dois tipos de funções que utilizaremos nessa classe. Tais funções são conhecidas na literatura por funções de variação regular e funções de variação lenta.

Considere a função $G : (0, \infty) \to (0, a), \quad a > 0$. Dizemos que G é uma *função de variação regular* se

$$\lim_{t\to\infty} \frac{G(tx)}{G(t)} = x^{\upsilon}, \quad \forall x > 0 \quad \text{e} \quad \upsilon \in \mathbb{R}, \tag{1.3}$$

onde υ é conhecido como índice de variação regular (*index of regular variation*). A notação para indicar que G é função de variação regular com índice υ será dada por $G \in RV_{\upsilon}$.

Se (1.3) for válido, podemos ainda escrever G como

$$G(x) = \ell_G(x)x^{\upsilon}, \quad \forall x > 0 \quad \text{e} \quad \upsilon \in \mathbb{R}, \tag{1.4}$$

onde $\ell_G(x)$ é conhecida por *função de variação lenta* (ℓ_G positiva) e satisfaz

$$\lim_{t\to\infty} \frac{\ell_G(tx)}{\ell_G(t)} = 1, \forall x > 0. \tag{1.5}$$

Para o caso de funções de variação lenta podemos utilizar a notação $\ell_G \in RV_0$, uma vez que equivale dizer que $\upsilon = 0$.

Definição 2 *Uma variável aleatória X, positiva, é dita ter distribuição com cauda de variação regular (com notação $X \in \mathcal{R}$), se sua função de sobrevivência puder ser escrita na forma*

$$1 - F(x) = \bar{F}(x) = \ell_F(x)x^{-\gamma}, \quad x > 0,$$

onde F corresponde à função de distribuição da variável aleatória, $\bar{F} = 1 - F$ é a função de sobrevivência, $\gamma > 0$ (desconhecido) é chamado de índice da cauda (tail index) e $\ell_F(x)$ é uma função de variação lenta.

É possível verificar que se $X \in \mathcal{R}$, então $X \in \mathcal{S}$, como pode ser observado em Goldie e Kluppelberg (1997) e Embrechts *et al.* (1997), onde também há mais detalhes sobre distribuições de caudas pesadas (além de outras classes).

A distribuição Fréchet é um exemplo de distribuição com cauda de variação lenta e a distribuição Weibull, com parâmetro de forma menor do que 1, é um exemplo de distribuição subexponencial que não pertence à classe de distribuições com cauda de variação regular.

1.3 Objetivos e organização

O intuito deste trabalho é comparar algumas formas de estimação intervalar para altos quantis oriundos de distribuições de caudas pesadas. Para isso, utilizamos três metodologias apresentadas em Peng e Qi (2006), que são denominadas por: aproximação pela normal, razão de verossimilhanças e *data tilting*. Além destas, fizemos também estimação intervalar, a partir do método delta, sob a suposição de que os dados são originários de uma distribuição Gama Generalizada, uma vez que esta, possui uma variedade considerável de distribuições conhecidas como casos particulares (dentre elas a distribuição Weibull), já que apresenta uma enorme flexibilidade na sua forma.

No Capítulo 2 apresentamos e detalhamos todas as metodologias utilizadas. No Capítulo 3 está dividido basicamente em três seções, onde apresentamos simulações de Monte Carlo utilizando as distribuições Weibull e Fréchet. Na Seção 3.1 utilizamos as distribuições de Weibull para verificar a convergência dos estimadores obtidos segundo o método delta para dados, supostamente, com distribuição Gama Generalizada. Na Seção 3.2 comparamos duas formas de se obter os limiares para os métodos da aproximação pela normal, razão de verossimilhanças e *data tilting*, no sentido de verificar quais destas formas – uma proposta por Chan *et al.* (2007), com base na experiência em trabalhos com o índice da cauda (*tail index*), e outra proposta por Beirlant *et al.* (2002), baseada no erro quadrático médio assintótico do estimador de Hill – fornecem os melhores intervalos de confiança. Na Seção 3.3 comparamos os quatros métodos utlizados neste trabalho, baseados nas probabilidades de cobertura, nas amplitudes médias e nos limites médios dos intervalos de confiança. No Capitulo 4 expomos um conjunto de dados reais referentes aos pagamentos de indenizações (em reais) de seguros de incêndios no ano de 2003 no Brasil, onde os métodos utilizados neste trabalho são aplicados e, são apresentados os intervalos de confiança para os quantis de ordem $0,99$ e $0,999$ dos pagamentos. Por fim, no Capítulo 5 apresentamos as conclusões obtidas neste trabalho e descrevemos algumas perspectivas.

Capítulo 2

Metodologias

Neste capítulo abordaremos quatro metodologias que serão utilizadas na estimação intervalar de altos quantis oriundos de distribuições de caudas pesadas, sendo estas: aproximação pela distribuição normal, razão de verossimilhanças, *data tilting* e gama generalizada. As três primeiras metodologias, que serão introduzidas a partir de agora, são apresentadas em Peng e Qi (2006).

Sejam $X_{1,}X_2, \ldots, X_n$ variáveis aleatórias independentes com mesma função de distribuição F, a qual pertence à classe de distribuições com cauda de variação regular exibida a seguir,

$$\bar{F}(x) = \ell_F(x)x^{-\gamma}, \quad x > 0, \tag{2.1}$$

onde $\bar{F}(x) = 1 - F(x) = \mathbb{P}(X > x)$, o índice da cauda (*tail index*), γ, é positivo e ℓ_F é uma função (positiva) de variação lenta, como apresentado em (1.4).

O $(1-p)$-ésimo quantil para a distribuição F será definido por $x_p = \sup\{x : F(x) \geq 1-p\}$. Mais especificamente, para o caso onde F possui função inversa, $x_p = \bar{F}^{-1}(p)$, onde $p = p_n \to 0$ quando $n \to \infty$ e $\bar{F}^{-1}(\cdot)$ corresponde à função inversa de $\bar{F}$.

Uma vez que $\ell_F(x)$ é uma função de variação lenta, podemos pensar que para valores grandes de x, $\ell_F(x) \approx c,\ c > 0$. Dessa forma temos que, para um determinado valor T (que será definido em breve) relativamente grande, a função de sobrevivência (ou probabilidade da cauda) em (2.1) pode ser escrita como

$$\bar{F}(x) = cx^{-\gamma}, \quad x > T. \tag{2.2}$$

Esta é uma maneira para que não seja necessário levar em consideração a forma da função ℓ_F, função esta que é desconhecida na prática. Além disso, o nosso interesse é estudar $\bar{F}(x)$ para altos valores de x $(x > T)$, pois queremos obter intervalos de confiança para altos quantis. Desse modo, basta criarmos uma nova variável $Y_i = \max\{X_i, T\}$. Logo, temos que a função de distribuição de Y_i será

$$F_{Y_i}(y) = \begin{cases} 0, & \text{se} \quad y < T; \\ \mathbb{P}(X_i \leq T), & \text{se} \quad y = T; \\ \mathbb{P}(X_i \leq y), & \text{se} \quad y > T. \end{cases} \tag{2.3}$$

Utilizando (2.2) temos que $\mathbb{P}(X_i \leq T) = F(T) = 1 - \bar{F}(T) = 1 - cT^{-\gamma}$ e, para $y > T$, $\mathbb{P}(X_i \leq y) = F(y) = 1 - \bar{F}(y) = 1 - cy^{-\gamma}$. Dessa forma, (2.3) fica escrita como

$$F_{Y_i}(y) = (1 - cT^{-\gamma}) \underset{\{T\}}{\mathbb{1}}(y) + (1 - cy^{-\gamma}) \underset{(T,\infty)}{\mathbb{1}}(y) \ , \tag{2.4}$$

onde $\underset{\{T\}}{\mathbb{1}}(y) = 1$ se $y = T$, e zero caso contrário. Analogamente, $\underset{(T,\infty)}{\mathbb{1}}(y) = 1$ se y pertencer ao intervalo (T, ∞), e zero caso contrário.

A função densidade de probabilidade (fdp) de Y_i será então

$$f_{Y_i}(y) = (1 - cT^{-\gamma}) \underset{\{T\}}{\mathbb{1}}(y) + c\gamma y^{-\gamma-1} \underset{(T,\infty)}{\mathbb{1}}(y) \ . \tag{2.5}$$

Note que Y_i é uma variável aleatória do tipo mista, com parte discreta no conjunto $\{T\}$ e parte absolutamente contínua no intervalo (T, ∞).

Como as variáveis $X_1, X_2, \ldots, X_n$ são independentes e identicamente distribuídas, então as variáveis $Y_1, Y_2, \ldots, Y_n$ também serão, já que Y_i é função mensurável de X_i, $i = 1, 2, \ldots, n$. Portanto, a função de verossimilhança para $\{(\delta_i, \max\{X_i, T\})\}_{i=1}^n$ será

$$L(\gamma, c) = \prod_{i=1}^{n} (c\gamma X_i^{-\gamma-1})^{\delta_i} (1 - cT^{-\gamma})^{1-\delta_i}, \tag{2.6}$$

com $\delta_i = \mathbb{1}(X_i > T)$, onde $\mathbb{1}(X_i > T) = 1$ se $X_i > T$, e zero caso contrário.

Sejam $X_{1,n} \leq X_{2,n} \leq \ldots \leq X_{n,n}$ as estatísticas de ordem de $X_1, X_2, \ldots, X_n$. Neste trabalho utilizaremos o valor de $T = X_{n-k,n}$, onde T é conhecido como limiar e os valores que pode assumir dependerão da escolha de k, que determina a fração amostral da cauda (*tail sample fraction*), k/n, e será comentado na Seção 2.1.

2.1 Aproximação pela distribuição normal

Comumente são feitas estimações baseadas na normalidade assintótica dos estimadores de máxima verossimilhança, sob algumas condições de regularidade, as quais não serão mencionadas neste tra-

balho, mas podem ser encontradas em, por exemplo, Casella e Berger (2001). Nesta seção faremos uso da normalidade assintótica do estimador do $(1-p)$-quantil, $\hat{x}_p$, obtido pelo método de máxima verossimilhança, como mencionaremos a seguir, para encontrar intervalos de confiança de x_p.

Utilizando (2.6) temos que a log-verossimilhança dos dados, $l(\gamma, c)$, será

$$l(\gamma, c) = k\log c + k\log\gamma - (\gamma+1)\sum_{i=1}^{k}\log X_{n-i+1,n} + (n-k)\log(1 - cX_{n-k,n}^{-\gamma}). \tag{2.7}$$

A partir de (2.7) teremos que os estimadores de máxima verossimilhança $(\hat{\gamma}_n, \hat{c}_n)$ para (γ, c) são (a demonstração encontra-se no Apêndice A.1)

$$\hat{c}_n = \frac{k}{n}X_{n-k,n}^{\hat{\gamma}_n}, \tag{2.8}$$

$$\hat{\gamma}_n = \left\{\frac{1}{k}\sum_{i=1}^{k}(\log X_{n-i+1,n} - \log X_{n-k,n})\right\}^{-1}, \tag{2.9}$$

em que $1/\hat{\gamma}_n$ é conhecido como estimador de Hill.

Proposto por Hill (1975), o estimador de Hill visa estimar $1/\gamma$, o qual é conhecido como índice do valor extremo (*extreme value index*) da distribuição. Haeusler e Teugels (1985) provam que tal estimador tem distribuição assintoticamente normal para seqüências intermediárias de valores de k, onde k é função de n e satisfaz

$$k \to \infty \quad \text{e} \quad \frac{k}{n} \to 0. \tag{2.10}$$

A escolha do valor de k tem implicações sobre o estimador de $1/\gamma$, pois k serve para definir a partir de qual valor, x, teremos $\mathbb{P}(X > x) = cx^{-\gamma}$, sendo, portanto, de fundamental importância. Contudo, tal escolha não é uma tarefa simples. Quando k cresce a variância do estimador de Hill diminui, mas seu viés aumenta (ver, por exemplo, Matthys *et al.* (2004) e Beirlant *et al.* (2004)). Algumas técnicas para a seleção da fração amostral serão utilizadas e comparadas no Capítulo 3.

Utilizando (2.2) teremos que um estimador natural para x_{p_n} é $\hat{x}_{p_n} = (p_n/\hat{c}_n)^{-1/\hat{\gamma}_n}$. Para derivar a normalidade assintótica de $\hat{x}_{p_n}$ é necessária uma condição mais estrita do que (2.1): suponha que exista uma função $A(t)$, tal que $A(t) \to 0$ quando $t \to \infty$, de tal forma que

$$\lim_{t\to\infty}\frac{U(tx)/U(t) - x^{1/\gamma}}{A(t)} = x^{1/\gamma}\frac{x^\rho - 1}{\rho}, \quad x > 0, \tag{2.11}$$

para algum $\rho < 0$, onde $U(x) = \left(\frac{1}{1-F}\right)^{-1}(x)$, ou seja, $U(x)$ corresponde à função inversa de

$\frac{1}{1-F(x)}$. A função $U(x)$ é conhecida como função quantil da cauda (*tail quantile function*). Assim (ver De Haan e Stadtmüller (1996)) $A(t)$ é uma função de variação regular com ρ sendo seu índice de variação regular (*index of regular variation*). Pode-se verificar que (2.11) implica (2.1) (ver Apêndice A.2).

O seguinte teorema pode ser derivado de Ferreira *et al.* (2003).

Teorema 2.1.1 (Peng e Qi, 2006) *Suponha que (2.10) e (2.11) sejam válidas. Se* $\sqrt{k}A(n/k) \to 0$, $np_n = O(k)$ *e* $\log(\frac{k}{np_n})/\sqrt{k} \to 0$ *quando* $n \to \infty$, *então*

$$\frac{\hat{\gamma}_n\sqrt{k}}{\log(k/(np_n))}\log\frac{\hat{x}_{p_n}}{x_{p_n}} \xrightarrow{D} N(0,1). \tag{2.12}$$

Dizer que $np_n = O(k)$ quando $n \to \infty$, significa dizer que $\frac{np_n}{k}$ é limitado quando $n \to \infty$. Em termos gerais podemos dizer que se tivermos duas seqüências $\{a_n\}_{n\in\mathbb{N}}$ e $\{b_n\}_{n\in\mathbb{N}}$ tais que $a_n = O(b_n)$, então existe um valor $M > 0$ onde $\lim_{n\to\infty}\left|\frac{a_n}{b_n}\right| = M$.

Portanto, baseado no limite do Teorema 2.1.1, um intervalo com nível de confiança α para x_{p_n} pode ser escrito como

$$I_\alpha^n = \left(\hat{x}_{p_n}\exp\left\{-z_\alpha\frac{\log\left(\frac{k}{np_n}\right)}{\hat{\gamma}_n\sqrt{k}}\right\}; \hat{x}_{p_n}\exp\left\{z_\alpha\frac{\log\left(\frac{k}{np_n}\right)}{\hat{\gamma}_n\sqrt{k}}\right\}\right),$$

sendo que z_α satisfaz $\mathbb{P}(|N(0,1)| \le z_\alpha) = \alpha$.

O Teorema 2.1.2 mostra que $\mathbb{P}(x_{p_n} \in I_\alpha^n) - \alpha = O((\log n)^{-2})$, se $\log(np_n) = O(\log n)$, garantindo que a precisão de cobertura para altos quantis, no geral, não é muito acurada.

Teorema 2.1.2 (Peng e Qi, 2006) *Sob as condições do Teorema 2.1.1,*

$$\begin{aligned}
&\mathbb{P}\left(\frac{\hat{\gamma}_n\sqrt{k}}{\log(k/(np_n))}\log\frac{\hat{x}_{p_n}}{x_{p_n}} \le x\right) - \Phi(x)\\
&= \frac{1}{3\sqrt{k}}\phi(x)(1+2x^2) - \phi(x)\frac{\gamma}{1-\rho}\sqrt{k}A(n/k) - \frac{1}{2}x\phi(x)\left(\log\frac{k}{np_n}\right)^{-2}\\
&\quad -o\left(\left(\log\frac{k}{np_n}\right)^{-2} + \frac{1}{\sqrt{k}} + \sqrt{k}|A(n/k)|\right)
\end{aligned}$$

uniformemente para $-\infty < x < \infty$, *onde* $\Phi(x)$ *e* $\phi(x)$ *correspondem às funções de distribuição e densidade de probabilidade da Normal padrão, respectivamente. Além do mais,*

$$\begin{aligned}\mathbb{P}(x_{p_n} \in I_\alpha^n) &= \alpha - z_\alpha \phi(z_\alpha)\left(\log\frac{k}{np_n}\right)^{-2} \\ &\quad + o\left(\left(\log\frac{k}{np_n}\right)^{-2} + \frac{1}{\sqrt{k}} + \sqrt{k}|A(n/k)|\right).\end{aligned}$$

2.2 Razão de verossimilhanças

O método exposto nesta seção é chamado de razão de verossimilhanças, e consiste, supondo que (2.2) seja válida, em estimar os parâmetros c e γ sob o seu espaço paramétrico e também estimá-los sob a restrição de que $x_{p_n} = (p_n/c)^{-1/\gamma}$, pois para $p_n = \bar{F}(x_{p_n})$, em (2.2), $p_n = cx_{p_n}^{-\gamma}$. A partir de então a estatística da razão de verossimilhanças será calculada.

Definamos $\hat{\gamma}_n$ e $\hat{c}_n$ como em (2.8) e (2.9), respectivamente, apresentados na Seção 2.1, ou seja, os estimadores de máxima verossimilhança para γ e c, respectivamente. Definamos l_1 como

$$l_1 = \max_{\gamma>0, c>0} l(\gamma, c) = l(\hat{\gamma}_n, \hat{c}_n).$$

Em seguida, maximizamos a log-verossimilhança em (2.7), $l(\gamma, c)$, sujeito a

$$\gamma > 0, \quad c > 0, \quad \gamma \log x_{p_n} + \log\left(\frac{p_n}{c}\right) = 0,$$

que será definida por $l_2(x_{p_n})$, isto é,

$$l_2(x_{p_n}) = \max\left\{l(\gamma, c) \,|\, \gamma > 0, c > 0, \gamma \log x_{p_n} + \log\left(\frac{p_n}{c}\right) = 0\right\}.$$

Note que $\{\gamma > 0 \quad \text{e} \quad c > 0\}$ correspondem ao espaço paramétrico e, $\gamma \log x_{p_n} + \log\left(\frac{p_n}{c}\right) = 0$ corresponde à restrição $x_{p_n} = (p_n/c)^{-1/\gamma}$.

Temos, assim, que $l_2(x_p) = l(\bar{\gamma}(\lambda), \bar{c}(\lambda))$, onde

$$\bar{\gamma}(\lambda) = \frac{k}{\sum_{i=1}^{k}(\log X_{n-i+1,n} - \log X_{n-k,n}) + \lambda \log X_{n-k,n} - \lambda \log x_p}, \tag{2.13}$$

$$\bar{c}(\lambda) = X_{n-k,n}^{\bar{\gamma}(\lambda)} \frac{k-\lambda}{n-\lambda}, \tag{2.14}$$

com λ satisfazendo

$$\bar{\gamma}(\lambda) \log x_{p_n} + \log\left(\frac{p_n}{\bar{c}(\lambda)}\right) = 0 \tag{2.15}$$

$$\bar{\gamma}(\lambda) > 0 \quad \text{e} \quad \lambda < k. \tag{2.16}$$

Temos que (2.13) e (2.14) são obtidos pelo método de multiplicadores de Lagrange (ver demonstração no Apêndice A.3).

Não é difícil verificar que (2.16) é equivalente a

$$\lambda < \min\left\{ \frac{\sum_{i=1}^{k}(\log X_{n-i+1,n} - \log X_{n-k,n})}{\log x_{p_n} - \log X_{n-k,n}}, k \right\}. \tag{2.17}$$

Dessa forma, temos que o logaritmo da razão de verossimilhanças multiplicado por menos dois (-2) será

$$l(x_p) = -2(l_2(x_p) - l_1). \tag{2.18}$$

Os intervalos de confiança poderão ser obtidos a partir de (2.18), com o uso do próximo teorema.

Teorema 2.2.1 (Peng e Qi, 2006) *Suponha que as condições do Teorema 2.1.1 sejam válidas. Então há uma única solução para (2.15) e (2.16), denotada* $\hat{\lambda}(x_p)$*, e ainda*

$$l(x_{p,0}) \xrightarrow{D} \chi^2_1, \tag{2.19}$$

com $\lambda = \hat{\lambda}(x_{p,0})$*, sendo* $x_{p,0}$ *o verdadeiro valor de* x_p*.*

Portanto, baseado no limite acima, um intervalo de confiança com nível α para x_p será

$$I^l_\alpha = \{x_p : l(x_p) \leq u_\alpha\} \tag{2.20}$$

em que u_α satisfaz $\mathbb{P}(\chi^2_1 \leq u_\alpha) = \alpha$. De acordo com Peng e Qi (2006), esta região de confiança tem probabilidade de cobertura assintoticamente correta (*asymptotically correct coverage probability*) e igual a α, isto é, $\mathbb{P}(x_p \in I^l_\alpha) \to \alpha$ quando $n \to \infty$.

Os limites de confiança para o método da razão de verossimilhanças são obtidos de forma aproximada, com base em (2.20). Devemos escolher um valor inicial do quantil x_p, de tal forma que o valor de $l(x_p)$, em (2.18), seja menor do que o α-ésimo quantil de uma distribuição χ^2 com 1 grau de liberdade (u_α). Em seguida, reduzimos (e aumentamos) o valor de x_p em passos pequenos, de tamanho $0,1$, por exemplo, até que $l(x_p) > u_\alpha$. Assim, o último valor de x_p, tal que $l(x_p) \leq u_\alpha$, será considerado o limite inferior (superior) do intervalo.

Note que o método da razão de verossimilhanças apresentado aqui é equivalente ao teste da razão de verossimilhanças comumente utilizado na literatura (algumas referências sobre o assunto

são Casella e Berger (2001), Dudewicz e Mishra (1988), Mood *et al.* (1974) e Schervish (1996)). A função $l(x_p)$ pode ser interpretada como a estatística do teste da razão de verossimilhanças aplicada às hipóteses $H_o : x_{p,0} = x_p$ contra $H_a : x_{p,0} \neq x_p$. Daí o motivo da restrição $\gamma \log x_p + \log\left(\frac{p_n}{c}\right) = 0$ para se estimar c e γ em $l_2(x_p)$. Logo, o intervalo de confiança deste método pode ser obtido com os valores, mínimo e máximo, que x_p pode assumir de tal forma que H_o não seja rejeitada.

2.3 Data tilting

O método *data tilting* foi proposto por Hall e Yao (2003), para séries temporais, e é baseado na verossimilhança empírica. O método da verossimilhança empírica, proposto inicialmente por Owen (1988), é uma abordagem não-paramétrica utilizada para a construção de regiões de confiança sem que seja necessário que se especifique a família de distribuições dos dados. Uma das vantagens deste método é que a forma da região de confiança, assim como o seu grau de assimetria, é determinada automaticamente pela amostra.

De modo geral podemos definir o método da verossimilhança empírica como a seguir.

Definição 3 *Sejam $X_1, X_2, \ldots, X_n$ vetores aleatórios independentes com mesma distribuição da variável $X \in \mathbb{R}^d$, a qual possui função de distribuição acumulada F_0. Teremos que os estimadores da verossimilhança empírica correspondem ao vetor de parâmetros $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq \mathbb{R}^p$, o qual maximiza a função R*

$$R(\boldsymbol{\theta}) = \max\left\{\prod_{i=1}^{n} n w_i | \sum_{i=1}^{n} w_i Z(X_i, \boldsymbol{\theta}) = 0, w_i \geq 0, \sum_{i=1}^{n} w_i = 1\right\}, \tag{2.21}$$

onde $\boldsymbol{\Theta}$ corresponde ao espaço paramétrico de $\boldsymbol{\theta}$, e $Z(X_i, \boldsymbol{\theta})$ são funções de X_i e $\boldsymbol{\theta}$ tais que $Z(X_i, \boldsymbol{\theta}) \in \mathbb{R}^s$ e $\mathbb{E}\left(Z(X_i, \boldsymbol{\theta})\right) = 0, \ \forall\ i = 1, 2, \ldots, n$.

Para o caso em que $\mathrm{Var}\left(Z(X_i, \boldsymbol{\theta})\right)$ é finita e com posto $q > 0$, tem-se que $-2\log(R(\boldsymbol{\theta})) \xrightarrow{D} \chi^2_q$ (ver Teorema 3.4 em Owen (2001), p. 41). Com este resultado é possível obter regiões de confiança para $\boldsymbol{\theta}$ sob o método da verossimilhança empírica.

Observe que, sob as restrições apontadas em (2.21), o vetor $\boldsymbol{w} = (w_1, w_2, \ldots, w_n)^\top$ que maximiza $\prod_{i=1}^{n} n w_i$ é o mesmo que maximiza $\sum_{i=1}^{n} \log w_i$. Dessa forma, utilizando multiplicadores de Lagrange, teremos que

$$w_i(\boldsymbol{\theta}) = \frac{1}{n(1 + \hat{\boldsymbol{\lambda}}^\top(\boldsymbol{\theta}) Z(X_i, \boldsymbol{\theta}))}, \ i = 1, 2, \ldots, n,$$

onde

$$\hat{\boldsymbol{\lambda}}(\boldsymbol{\theta}) = \arg\min_{\boldsymbol{\lambda}\in\mathbb{R}^s} \left\{ -\sum_{i=1}^{n} \log\left(1 + \boldsymbol{\lambda}^\top Z(X_i, \boldsymbol{\theta})\right) \right\}.$$

Mais informações sobre $w_i(\boldsymbol{\theta})$ e $\hat{\boldsymbol{\lambda}}(\boldsymbol{\theta})$, além de mais detalhes sobre verossimilhança empírica, podem ser consultados em Kitamura (2006).

A título de ilustração, para o caso $\theta = \mu$, com μ sendo a média populacional dos dados, teremos que $Z(X, \mu) = X - \mu$. Se $\theta = x_p$, com x_p sendo o $(1-p)$-quantil de X, $0 < p < 1$, teremos que $Z(X, x_p) = \mathbb{1}(X \leq x_p) - (1-p)$, onde $\mathbb{1}(X \leq x_p) = 1$ se $X \leq x_p$, e zero caso contrário. Mais exemplos e outros detalhes podem ser obtidos, de forma mais completa, em Owen (2001).

De acordo com Chan *et al.* (2007), o método *data tilting* é uma generalização do método da verossimilhança empírica, que não só possui todas as suas boas propriedades, como também admite uma grande variedade de funções de distância (*distance functions*). Neste trabalho, assim como em Hall e Yao (2003) e Peng e Qi (2006), trabalharemos com a chamada função de distância do tipo potência (*power distance function*).

Inicialmente, para alguns pesos fixos $\boldsymbol{q} = (q_1, q_2, \ldots, q_n)$, tais que $q_i \geq 0, \forall i = 1, 2, \ldots, n$ e $\sum_{i=1}^{n} q_i = 1$, obtém-se

$$(\hat{\gamma}(\boldsymbol{q}), \hat{c}(\boldsymbol{q})) = \arg\max_{(\gamma,c)} \sum_{i=1}^{n} q_i \log\left\{ (c\gamma X_i^{-\gamma-1})^{\delta_i} (1 - cX_{n,n-k}^{-\gamma})^{1-\delta_i} \right\},$$

o que resulta em (ver Apêndice A.4)

$$\begin{aligned} \hat{\gamma}(\boldsymbol{q}) &= \frac{\sum_{i=1}^{n} q_i \delta_i}{\sum_{i=1}^{n} q_i \delta_i (\log X_i - \log X_{n,n-k})}, \\ \hat{c}(\boldsymbol{q}) &= X_{n,n-k}^{\hat{\gamma}(\boldsymbol{q})} \sum_{i=1}^{n} q_i \delta_i, \end{aligned}$$

onde o vetor de pesos $\boldsymbol{q}$ é obtido minimizando

$$D_{\rho_0}(\boldsymbol{q}) = \begin{cases} (\rho_0(1-\rho_0))^{-1} \left\{ 1 - n^{-1} \sum_{i=1}^{n} (nq_i)^{\rho_0} \right\}, & \text{se} \quad \rho_0 \neq 0, 1; \\ -n^{-1} \sum_{i=1}^{n} \log(nq_i), & \text{se} \quad \rho_0 = 0; \\ \sum_{i=1}^{n} q_i \log(nq_i), & \text{se} \quad \rho_0 = 1. \end{cases}$$

A função $D_{\rho_0}(\boldsymbol{q})$ é uma medida de distância entre $\boldsymbol{q}$ e o vetor uniforme, isto é, $q_i = \frac{1}{n}$ $(i =$

$1, 2, \ldots, n$). Então, $\boldsymbol{q}$ é escolhido de tal forma que a distância $D_{\rho_0}(\boldsymbol{q})$ seja mínima. Mais especificamente, resolve-se $(2n)^{-1}L(x_p) = \min_q D_{\rho_0}(\boldsymbol{q})$ sujeita às restrições

$$q_i \geq 0, \quad \sum_{i=1}^{n} q_i = 1, \quad \hat{\gamma}(\boldsymbol{q}) \log \frac{x_p}{X_{n,n-k}} = \log \frac{\sum_{i=1}^{n} q_i \delta_i}{p_n}.$$

Observe que $\hat{\gamma}(\boldsymbol{q}) \log(x_{p_n}/X_{n,n-k}) = \log(\sum_{i=1}^{n} q_i \delta_i / p_n)$ é equivalente a $x_{p_n} = (p_n/\hat{c}(\boldsymbol{q}))^{-1/\hat{\gamma}(\boldsymbol{q})}$.

Peng e Qi (2006) alegam que o caso $\rho_0 = 1$ possui boas propriedades de robustez. Por este motivo, trabalharemos aqui com o mesmo caso.

Sejam

$$A_1(\lambda_1) = 1 - \frac{n-k}{n} e^{-1-\lambda_1} \text{ e } A_2(\lambda_1) = A_1(\lambda_1) \frac{\log(x_{p_n}/X_{n,n-k})}{\log(A_1(\lambda_1)/p_n)}.$$

Assim, pelo método dos Multiplicadores de Lagrange, teremos (ver Apêndice A.5)

$$\begin{aligned} q_i &= q_i(\lambda_1, \lambda_2) \\ &= \begin{cases} \frac{1}{n} e^{-1-\lambda_1}, & \text{se } \delta_i = 0, \\ \frac{1}{n} \exp\Big\{ -1 - \lambda_1 \\ \qquad + \lambda_2 \Big(\frac{\log(x_{p_n}/X_{n,n-k})}{A_2(\lambda_1)} - \frac{1}{A_1(\lambda_1)} \\ \qquad - \frac{A_1(\lambda_1) \log(X_i/X_{n,n-k}) \log(x_{p_n}/X_{n,n-k})}{A_2^2(\lambda_1)} \Big) \Big\}, & \text{se } \delta_i = 1, \end{cases} \end{aligned} \tag{2.22}$$

onde λ_1 e λ_2 satisfazem

$$\sum_{i=1}^{n} q_i = 1, \quad \hat{\gamma}(\boldsymbol{q}) \log \frac{x_{p_n}}{X_{n,n-k}} = \log \frac{\sum_{i=1}^{n} q_i \delta_i}{p_n}. \tag{2.23}$$

O próximo teorema garante solução para as equações apresentadas em (2.23).

Teorema 2.3.1 (Peng e Qi, 2006) *Suponha que as condições no Teorema 2.1.1 sejam válidas. Então, com probabilidade tendendo a 1, existe solução para (2.23), denominadas por* $(\hat{\lambda}_1(x_p), \hat{\lambda}_2(x_p))$,

tais que, para $(\lambda_1, \lambda_2) = (\hat{\lambda}_1(x_p), \hat{\lambda}_2(x_p))$,

$$\begin{aligned} -\log\left(1 + \frac{\sqrt{k}\sqrt{\log(k/(np_n))}}{n-k}\right) &\leq 1 + \lambda_1 \\ &\leq -\log\left(1 - \frac{\sqrt{k}\sqrt{\log(k/(np_n))}}{n-k}\right), \end{aligned} \tag{2.24}$$

$$|\lambda_2| \leq k^{-1/4}\frac{k/n}{\log(k/(np_n))} \tag{2.25}$$

e $L(x_{p,0}) \xrightarrow{D} \chi^2_1$ *com* $(\lambda_1, \lambda_2) = (\hat{\lambda}_1(x_{p,0}), \hat{\lambda}_2(x_{p,0}))$ *na definição de* $L(x_{p,0})$*, onde* $x_{p,0}$ *corresponde ao verdadeiro valor de* x_p*.*

Portanto, baseado no limite acima, uma região de confiança com nível α para x_p é

$$I^t_\alpha = \{x_p : L(x_p) \leq u_\alpha\}, \tag{2.26}$$

onde u_α satisfaz $\mathbb{P}(\chi^2_1 \leq u_\alpha) = \alpha$. De acordo com Peng e Qi (2006), esta região de confiança tem probabilidade de cobertura assintoticamente correta (*asymptotically correct coverage probability*) e igual a α, isto é, $\mathbb{P}(x_p \in I^t_\alpha) \to \alpha$ quando $n \to \infty$.

Os limites de confiança para o método do *data tilting* são obtidos, assim como no método da razão de verossimilhanças, de forma aproximada, com base em $L(x_p)$. Devemos escolher um valor inicial do quantil x_p, de tal forma que o valor de $L(x_p)$ seja menor do que o α-ésimo quantil de uma distribuição χ^2 com 1 grau de liberdade (u_α). Em seguida, reduzimos (e aumentamos) o valor de x_p em passos pequenos, de tamanho $0,1$, por exemplo, até que $L(x_p) > u_\alpha$. Assim, o último valor de x_p, onde $L(x_p) \leq u_\alpha$, será considerado o limite inferior (superior) do intervalo.

2.4 Gama generalizada

A distribuição Gama é comumente utilizada para prever eventos extremos, principalmente nas áreas de hidrologia e ciências atuariais. Melo (2006), por exemplo, utiliza a distribuição Gama como alternativa para modelar a distribuição da severidade (valor) de sinistros acima de um determinado limite de retenção e, verifica que tal distribuição, comparada a outras três (a saber: Log-Normal, Pareto Generalizada e Pareto Generalizada Modificada), apresenta melhor desempenho para limites baixos de retenção.

Outra distribuição muito utilizada na literatura corresponde à distribuição Weibull. No ramo de seguros, esta distribuição é muito utilizada para modelar sinistros (mais detalhes em Embrechts *et al.* (1997)).

As distribuições Gama e Weibull, podem ser obtidas como casos particulares da Gama Generalizada, que possui um segundo parâmetro de forma, β, o que faz com que sua distribuição seja mais flexível do que as duas distribuições supracitadas. Ashkar e Ouarda (1998) propõem intervalos de confiança aproximados para quantis da Gama Generalizada, os quais se mostraram úteis em aplicações na hidrologia, em conjuntos de dados pequenos. Van Noortwijk (1999) faz inferências para os quantis da Gama Generalizada utilizando uma abordagem bayesiana, e aplica tal método para obter o nível de retorno de 1250 anos no conjunto de dados dos máximos anuais do rio Reno (*Rhine*), em Lobith, na Holanda, de 1901 a 1995.

Apesar da aplicabilidade da distribuição Gama Generalizada, não é comum o uso desta distribuição na área de ciências atuariais. Nossa proposta aqui é de calcular intervalos de confiança, utilizando o método delta, para os quantis da Gama Generalizada, e então verificar se a mesma é apropriada para modelar valores de sinistros.

Proposta por Stacy (1962), a distribuição gama generalizada, com função densidade de probabilidade dada em (2.27), pode assumir uma variedade de distribuições diferentes como casos particulares, como podemos observar na Tabela 2.1, apresentada em Stacy e Mihram (1965), com $\beta > 0$, $\zeta > 0$ e $\alpha > 0$, e χ^2_n e χ_n correspondem às distribuições Qui-quadrado e Qui, respectivamente, ambas com n graus de liberdade.

Distribuição	β	ζ	α
Exponencial	1	1	α
Gama	1	ζ	α
Weibull	β	1	α
χ^2_n	1	$n/2$	2
χ_n	2	$n/2$	$\sqrt{2}$
Half-Normal	2	$1/2$	$\sqrt{2}$
Normal Circular	2	1	$\sqrt{2}$
Normal Esférica	2	$3/2$	$\sqrt{2}$
Rayleigh	2	1	$c\sqrt{2}$

Tabela 2.1 *Distribuições que podem ser escritas como casos particulares da Gama Generalizada apresentadas em Stacy e Mihram (1965)*

Para uma melhor compreensão a respeito da distribuição Gama Generalizada vamos definí-la melhor a seguir.

Definição 4 *Seja X uma variável aleatória com distribuição Gama Generalizada de parâmetros β, ζ e α ($X \sim GG(\beta, \zeta, \alpha)$), com $\beta > 0$, $\zeta > 0$, $\alpha > 0$. Temos que a função densidade de probabilidade de X será*

$$f_X(x) = \frac{\beta x^{\beta\zeta - 1}}{\Gamma(\zeta)\alpha^{\beta\zeta}} \exp\left\{-\left(\frac{x}{\alpha}\right)^{\beta}\right\} \mathbb{1}_{(0,\infty)}(x) \ , \tag{2.27}$$

onde α corresponde ao parâmetro de escala e, β e ζ representam os parâmetros de forma da distribuição.

Assim, se tivermos $X_1, X_2, \ldots, X_n$, independentes com distribuição Gama Generalizada, $\mathrm{GG}(\beta, \zeta, \alpha)$, então a função de log-verossimilhança dos dados será

$$l(\beta, \zeta, \alpha) = n \log \beta - n\zeta\beta \log \alpha - n \log(\Gamma(\zeta)) + (\beta\zeta - 1) \sum_{i=1}^{n} \log x_i - \sum_{i=1}^{n} \left(\frac{x_i}{\alpha}\right)^{\beta}. \tag{2.28}$$

Os estimadores de máxima verossimilhança, $\hat{\beta}, \hat{\zeta}$ e $\hat{\alpha}$, são estimados iterativamente. Neste trabalho, os intervalos de confiança foram calculados com a utilização do software R. A maximização dos parâmetros foi feita com base no método de Nelder e Mead (1965), a partir do comando `optim` (ver R Development Core Team (2008) para mais detalhes).

A partir das estimativas dos parâmetros obtemos a estimativa do $(1-p)$-ésimo quantil, $\hat{x}_p$. O $(1-p)$-ésimo quantil da Gama Generalizada, x_p, pode ser obtido a partir do $(1-p)$-ésimo quantil da distribuição Gama , y_p. Seja Y uma variável aleatória com distribuição Gama de parâmetros ζ e α ($Y \sim \mathrm{Gama}(\zeta, \alpha)$), com o parâmetro de forma $\zeta > 0$ e o parâmetro de escala $\alpha > 0$. Como mostrado na Tabela 2.1, Y é um caso particular da Gama Generalizada, quando $\beta = 1$. É possível também escrever X como função de Y, e vice-versa, sob a seguinte relação

$$X = \alpha \left(\frac{Y}{\alpha}\right)^{\frac{1}{\beta}}, \tag{2.29}$$

e então verificar que o $(1-p)$-quantil da distribuição Gama Generalizada, x_p, pode ser obtido a partir do $(1-p)$-quantil da Gama, y_p (ver Apêndice A.6).

Os intervalos de confiança para os quantis são obtidos através do método Delta, o qual definiremos a seguir.

Seja $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^\top$ um vetor de parâmetros a ser estimado. Denote por $\hat{\boldsymbol{\theta}}$ o estimador de máxima verossimilhança de $\boldsymbol{\theta}$. Em geral, sob condições de regularidade necessários para a obtenção dos resultados assintóticos dos estimadores de máxima verossimilhança (mais detalhes em Lehmann

e Casella (2003))

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{D} N(0, \mathrm{Var}(\hat{\boldsymbol{\theta}}))$$

então, para g uma função contínua, temos que

$$\sqrt{n}(g(\hat{\boldsymbol{\theta}}) - g(\boldsymbol{\theta})) \xrightarrow{D} N(0, [g'(\boldsymbol{\theta})]^\top \mathrm{Var}(\hat{\boldsymbol{\theta}}) g'(\boldsymbol{\theta}))$$

com $g'(\boldsymbol{\theta}) = \dfrac{\partial g(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$.

No nosso caso, $\boldsymbol{\theta} = (\beta, \zeta, \alpha)^\top$ e $g : \mathbb{R}^3_+ \to \mathbb{R}$ corresponde à inversa da função de distribuição da Gama Generalizada aplicada aos parâmetros, com uma dada probabilidade $(1 - p)$. Os quantis foram calculados utilizando o pacote `VGAM` (mais detalhes em Yee (2008)). As derivadas parciais em $g'(\boldsymbol{\theta})$ foram calculadas numericamente, utilizando Toomet e Henningsen (2008).

Capítulo 3

Simulações

Com o intuito de avaliar o desempenho dos quatro métodos utilizados neste trabalho, faremos simulações de Monte Carlo com duas distribuições diferentes, uma com cauda de variação regular e outra do tipo subexponencial (sem que a mesma possua cauda de variação regular), sendo estas, respectivamente: Fréchet(α), com função de distribuição acumulada $F(x) = \exp\left(-x^{-\alpha}\right),\ x > 0$, e Weibull($\beta, \alpha$), com função de distribuição acumulada $F(x) = 1 - \exp\left\{-\left(\frac{x}{\alpha}\right)^{\beta}\right\},\ x > 0$. Gostaríamos de lembrar que a distribuição Weibull só será do tipo subexponencial se $0 < \beta < 1$.

Geramos 10000 amostras de tamanhos 1000, 1500, 2000 e 2500 com distribuição Weibull de parâmetros $\beta = 0,3$ e $\alpha = 1$, cujo gráfico da função densidade de probabilidade pode ser encontrado na Figura 3.1, e mais 10000 amostras de tamanhos 1000 e 2000 com distribuição Fréchet de parâmetro $\alpha = 1$, cujo gráfico da função densidade de probabilidade é exibido na Figura 3.2. Observamos, nas duas figuras, que ambas as distribuições possuem formas parecidas.

Este capítulo encontra-se dividido como segue. Na Seção 3.1 verificamos a convergência dos estimadores obtidos, segundo o método delta (gama generalizada). Nesta seção utilizamos 10000 amostras de tamanhos 1000, 1500, 2000 e 2500 com distribuição Weibull, que é caso particular da distribuição Gama Generalizada. Na Seção 3.2 comparamos o valor $k = 1,5(\log n)^2$, proposto por Chan *et al.* (2007), com o valor k obtido de tal forma que o erro quadrático médio assintótico do estimador de Hill seja mínimo (ver Apêndice B), proposto por Beirlant *et al.* (2002), no sentido de verificar qual destes fornece melhores resultados no cálculo dos intervalos de confiança, com probabilidades de cobertura mais próximas do nível de confiança e amplitudes médias menores, em relação aos métodos da aproximação pela normal, razão de verossimilhanças e *data tilting*. Na Seção 3.2 foram utilizadas as amostras de tamanho 1000 e 2000 das distribuições Weibull e Fréchet. A distribuição Weibull não possui uma condição necessária para os três métodos utilizados (ser distribuição com cauda de variação regular), mas será utilizada com a finalidade de verificar qual a escolha de k que deixa os métodos mais robustos com relação à quebra de tal suposição. Na Seção 3.3 comparamos todos os métodos utilizados nesta dissertação, com o intuito de verificar

qual destes é mais apropriado, no sentido de fornecer intervalos de confiança com probabilidades de cobertura mais próximas do nível de confiança e amplitudes médias menores.

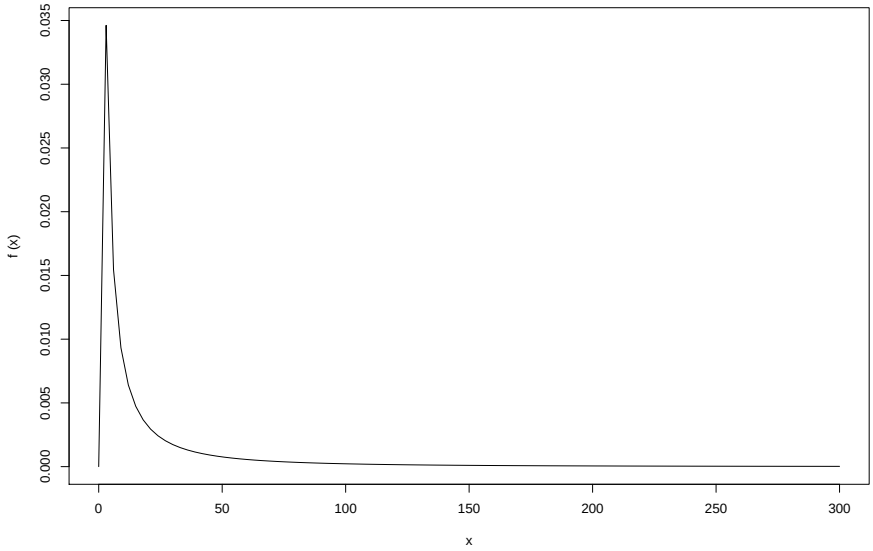

Figura 3.1 *Função densidade de probabilidade da distribuição Weibull com parâmetros* $\beta = 0,3$ *e* $\alpha = 1$

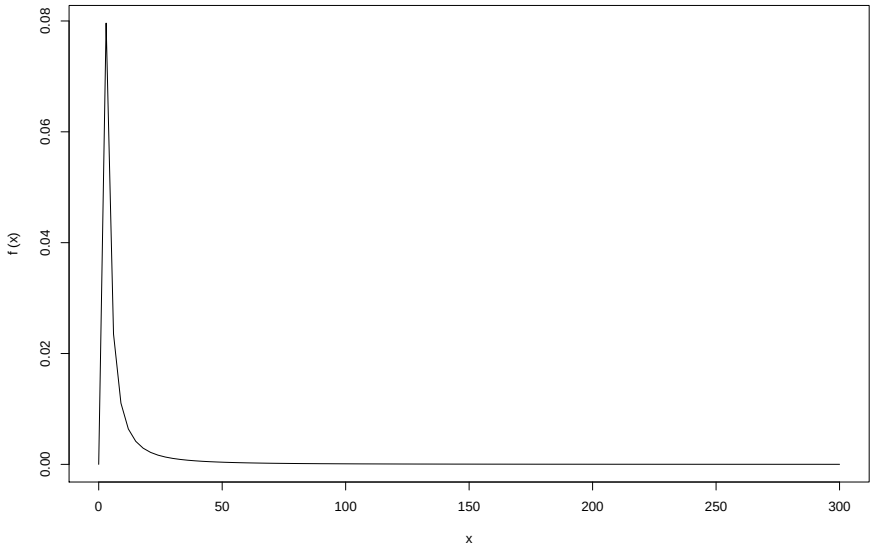

Figura 3.2 *Função densidade de probabilidade da distribuição Fréchet com parâmetro* $\alpha = 1$

3.1 Convergência utilizando o método delta

Utilizando 10000 amostras geradas com distribuição Weibull, tentamos comparar a proximidade entre as distribuições amostrais dos estimadores e a distribuição Normal.

As estimativas foram obtidas maximizando, iterativamente, a função log-verossimilhança (2.28), como explicado no Capítulo 2.4. Sob esta configuração de parâmetros, houve sempre convergência no processo de estimação. Contudo, em algumas amostras, a estimativa da matriz de informação de Fisher não resultou em uma matriz positiva definida (exibimos a contagem dessas amostras na Tabela 3.1). Desta forma, foram levadas em consideração apenas as amostras em que tais estimativas resultaram em matrizes positivas definidas, já que toda matriz de variâncias e covariâncias deve ser deste tipo.

Nas Figuras 3.3, 3.4 e 3.5 e na Tabela 3.2 podemos observar bons indícios de que a distribuição amostral das estimativas dos parâmetros da distribuição Gama Generalizada está convergindo para a distribuição Normal, cujas média e variância foram estimadas a partir das estimativas das réplicas de Monte Carlo. Na Tabela 3.2 vemos que a média das estimativas está convergindo para os verdadeiros valores de seus respectivos parâmetros e, ainda, o erro quadrático médio (EQM) estimado diminui, à medida que o tamanho amostral aumenta.

Para cada réplica, utilizando as estimativas dos seus respectivos parâmetros, estimamos os quantis de ordem $0,99$ e $0,999$. Dessas estimativas observamos (ver Figuras 3.6 e 3.7 e Tabela 3.3) que, com o aumento do tamanho amostral, o quantil médio estimado vai se aproximando do verdadeiro valor do quantil, os quais correspondem a 162,4871 e 627,7545, respectivamente.

Um fato que achamos interessante mencionar é que, apesar das estimativas dos parâmetros nas amostras de tamanho 1000 não apresentarem um comportamento muito parecido com o de uma distribuição Normal, as estimativas dos quantis de ordem $0,99$ e $0,999$, para o mesmo tamanho amostral, já mostraram um comportamento parecido com o comportamento gaussiano.

As probabilidades de cobertura e amplitudes médias encontram-se na Tabela 3.4, onde observamos que, quando o tamanho amostral cresce, a probabilidade de cobertura dos intervalos aparenta convergir para o nível de confiança utilizado aqui (0,90) e, a amplitude média dos mesmos diminui. Como as probabilidades de cobertura mostraram-se maiores do que o nível de confiança (0,90) para todos os tamanhos amostrais utilizados, suspeitamos que a probabilidade de cobertura dos intervalos de confiança está tendendo para 0,90 pela direita, à medida que o tamanho amostral au-

menta, ou seja, temos bons indícios de que, para amostras grandes, a probabilidade dos intervalos de confiança conterem o verdadeiro valor do quantil será, pelo menos, o nível de confiança.

n	Casos
1000	9
1500	6
2000	1
2500	4

Tabela 3.1 *Número de casos onde a estimativa da matriz de informação não foi positiva definida, de acordo com o tamanho amostral, em um total de 10000 amostras.*

n	$\hat{\beta}$	$\hat{\zeta}$	$\hat{\alpha}$	EQM($\hat{\beta}$)	EQM($\hat{\zeta}$)	EQM($\hat{\alpha}$)
1000	0,30301	1,01132	1,20051	0,00098	0,03026	0,62929
1500	0,30282	1,00279	1,15435	0,00066	0,01911	0,40142
2000	0,30208	1,00210	1,11761	0,00049	0,01429	0,28508
2500	0,30216	0,99954	1,10701	0,00041	0,01164	0,23650

Tabela 3.2 *Estimativas médias dos parâmetros da distribuição Gama Generalizada, para cada tamanho amostral, com $\beta = 0,3$, $\zeta = 1$ e $\alpha = 1$.*

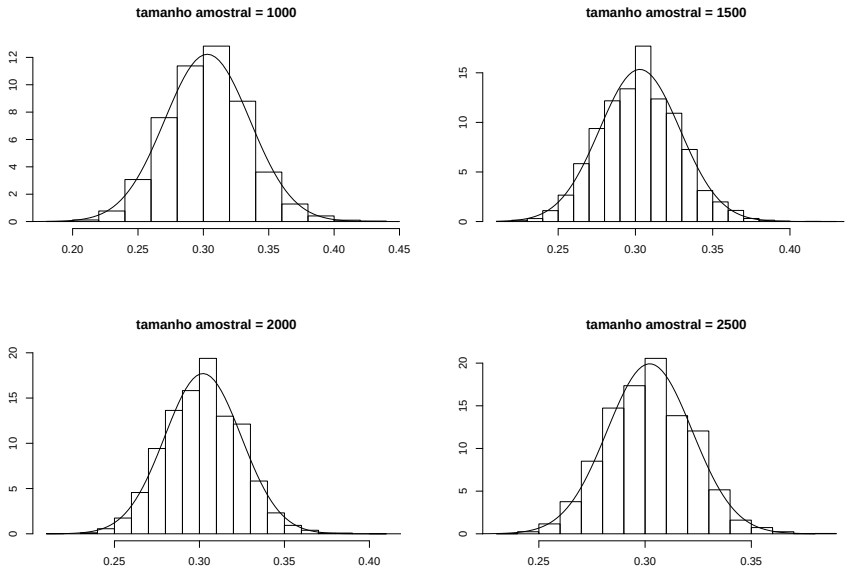

Figura 3.3 *Histogramas das 10000 estimativas de β comparadas com a distribuição normal*

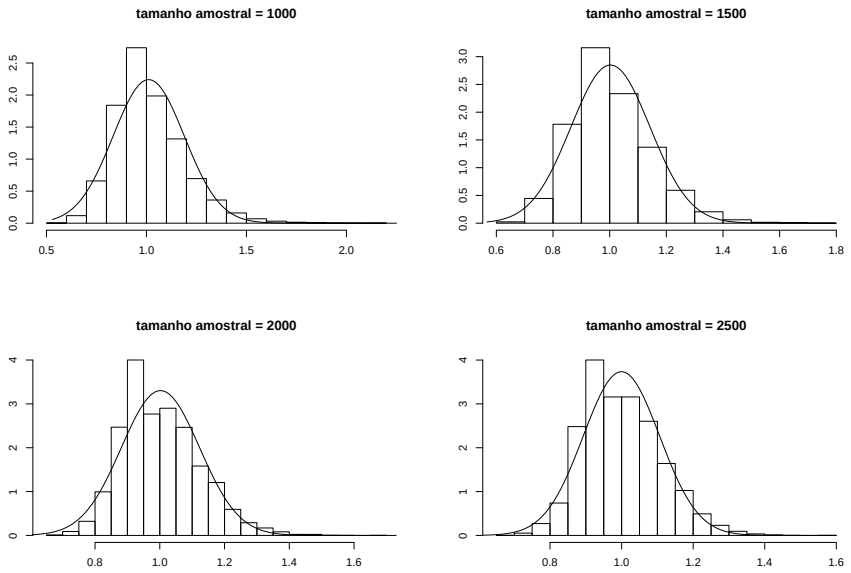

Figura 3.4 *Histogramas das 10000 estimativas de ζ comparadas com a distribuição normal*

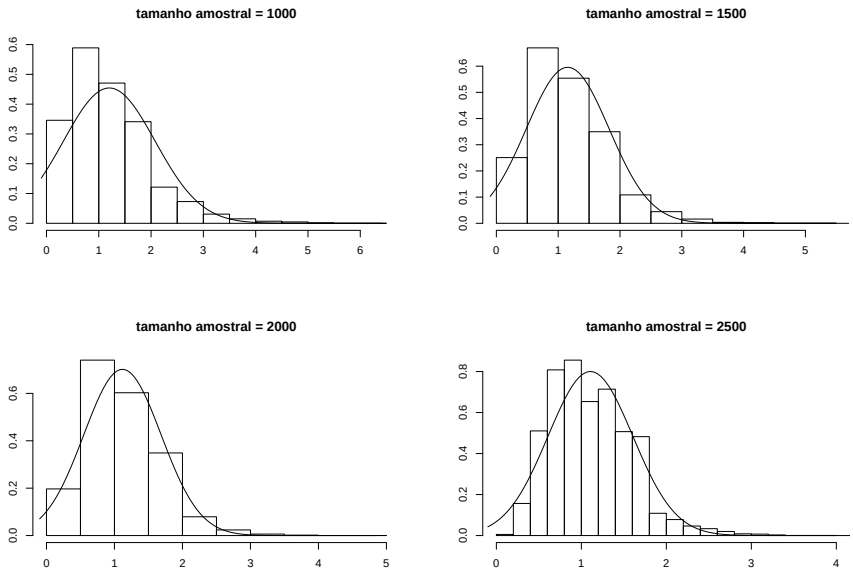

Figura 3.5 *Histogramas das 10000 estimativas de α comparadas com a distribuição normal*

3.2 Escolha do limiar

O valor de k nos auxilia na escolha do limiar T (sendo este a $(n-k)$-ésima estatística de ordem, ou seja, $T = X_{n-k,n}$), a partir do qual a cauda da distribuição dos dados deve se comportar como na forma (2.2).

Como já mencionado anteriormente, Chan *et al.* (2007) propuseram, com base em suas ex-

	$q_{0,99}$		$q_{0,999}$	
n	$\hat{q}_{0,99}$	EQM($\hat{q}_{0,999}$)	$\hat{q}_{0,999}$	EQM($\hat{q}_{0,999}$)
1000	163,0447	739,0026	635,6514	22150,6800
1500	162,5643	498,0292	630,8571	14609,9500
2000	162,7204	364,6889	630,7032	10638,1000
2500	162,3698	296,4878	628,2951	8671,5860

Tabela 3.3 *Estimativas médias para os quantis de ordem* $0,99$ *e* $0,999$*, para 10000 amostras geradas da distribuição Weibull, variando o tamanho amostral, segundo o método delta (gama generalizada), com* $q_{0,99} = 162,4871$ *e* $q_{0,999} = 627,7545$*.*

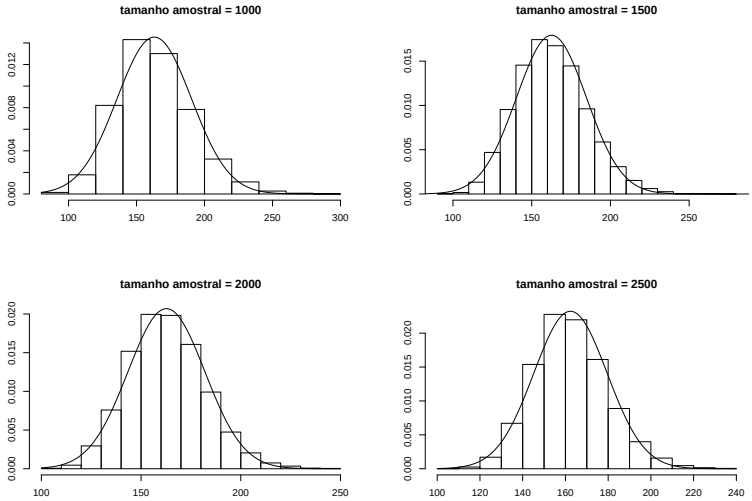

Figura 3.6 *Histogramas das 10000 estimativas do quantil de ordem* $0,99$ *comparadas com a distribuição normal*

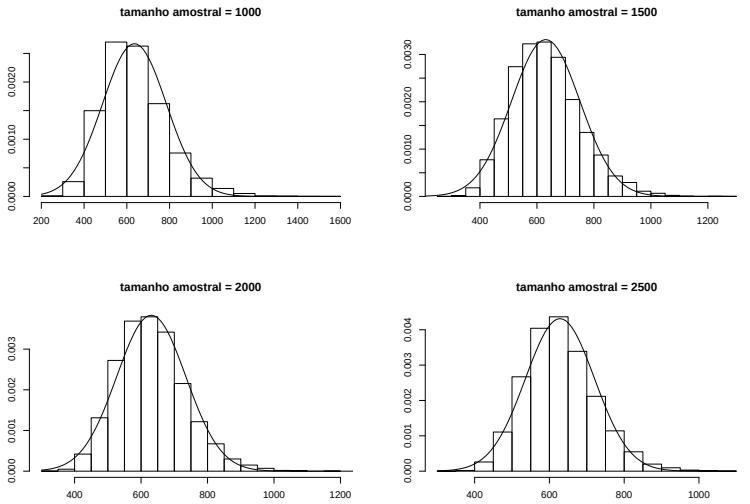

Figura 3.7 *Histogramas das 10000 estimativas do quantil de ordem* $0,999$ *comparadas com a distribuição normal*

	$q_{0,99}$		$q_{0,999}$	
n	Cobert.	Amplitude	Cobert.	Amplitude
1000	0,9108	90,2123	0,9153	491,5739
1500	0,9018	73,1634	0,9107	395,9505
2000	0,9049	63,3943	0,9096	342,4411
2500	0,9023	56,4704	0,9014	304,2357

Tabela 3.4 *Probabilidades de cobertura e amplitudes médias dos intervalos de 90% de confiança para os quantis de ordem* $0,99$ *e* $0,999$*, para 10000 amostras geradas da distribuição Weibull, variando o tamanho amostral, segundo o método delta (gama generalizada), com* $q_{0,99} = 162,4871$ *e* $q_{0,999} = 627,7545$*.*

	$q_{0,99}$		$q_{0,999}$	
n	Lim. Inf.	Lim. Sup.	Lim. Inf.	Lim. Sup.
1000	117,9385	208,1508	389,8645	881,4383
1500	125,9825	199,1460	432,8819	828,8324
2000	131,0232	194,4175	459,4826	801,9237
2500	134,1345	190,6050	476,1772	780,4129

Tabela 3.5 *Limites inferior e superior médios dos intervalos de 90% de confiança para os quantis de ordem* $0,99$ *e* $0,999$*, para 10000 amostras geradas da distribuição Weibull, variando o tamanho amostral, segundo o método delta (gama generalizada), com* $q_{0,99} = 162,4871$ *e* $q_{0,999} = 627,7545$*.*

periências com o índice da cauda (*tail index*), um valor para o limiar, com base em $k = 1,5\,(\log n)^2$ (onde n corresponde ao tamanho amostral), a ser utilizado nos métodos da aproximação pela normal, razão de verossimilhanças e *data tilting*. Dessa forma, resolvemos verificar se as probabilidades de cobertura dos intervalos apresentam valores próximos do nível de confiança, além de observar as amplitudes médias dos intervalos e comparar tal escolha com uma outra, comumente utilizada, proposta por Beirlant *et al.* (2002), na qual o valor de k é obtido de tal forma que o erro quadrático médio assintótico do estimador de Hill seja mínimo. Para isso, calculamos os intervalos com nível de 90% de confiança para os quantis de ordem $0,99$ e $0,999$, sob os três métodos de estimação intervalar supracitados. Nesta seção trabalhamos apenas com as simulações de tamanhos amostrais 1000 e 2000 das distribuições Weibull e Fréchet.

Com o intuito de simplificar, denotaremos a escolha de $k = 1,5(\log n)^2$ por k_1 e o valor de k que minimiza o erro quadrático médio assintótico por k_2.

Inicialmente, gostaríamos de destacar que, o valor k_1 é sempre o mesmo, dependendo apenas do tamanho amostral. No nosso caso, temos que k_1 corresponde a 71 e 86 para amostras de tamanhos 1000 e 2000, respectivamente. Contudo, como pode ser observado no Apêndice B, o valor de k que

minimiza o erro quadrático médio assintótico do estimador de Hill, k_2, depende não só do tamanho amostral, n, como também da própria amostra, variando portanto para cada conjunto de dados. Para amostras de tamanho 1000, procuramos encontrar k_2 entre os valores 20 e 300, e para amostras de tamanho 2000, procuramos encontrar k_2 entre os valores 30 e 500. O intuito disso é simplesmente não obter valores muito baixos, que resultariam em um viés elevado para o estimador de Hill, ou valores muito altos, que implicariam uma variância muito grande. Na Tabela 3.6 podemos observar algumas estatísticas descritivas para os valores k_2, para as amostras de tamanhos 1000 e 2000, geradas a partir das distribuições Weibull e Fréchet. Observa-se que o valor médio de k_2 para a distribuição Weibull é bem inferior ao valor médio de k_2 para a distribuição Fréchet, além de indícios de uma variabilidade menor para a distribuição Weibull.

Distribuição	n	Min.	1^o Quartil	Mediana	Média	3^o Quartil	Max.	D. padrão
Weibull	1000	20,00	25,00	38,00	53,35	61,00	300,00	51,58
	2000	30,00	38,00	56,00	71,21	86,00	500,00	55,43
Fréchet	1000	20,00	171,00	229,00	214,90	274,00	300,00	69,07
	2000	30,00	228,00	328,00	316,00	414,00	500,00	119,85

Tabela 3.6 *Estatísticas descritivas (mínimo, primeiro quartil, mediana, média, terceiro quartil, máximo e desvio padrão) dos valores de k que minimizam o erro quadrático médio assintótico do estimador de Hill, k_2, para 10000 amostras de tamanhos 1000 e 2000 das distribuições Weibull e Fréchet.*

Como o nosso intuito nesta seção é verificar qual a forma de escolha do valor de k que fornece melhores intervalos de confiança, vamos analisar, para os métodos aproximação pela normal, razão de verossimilhança e *data tilting*, como os intervalos de confiança para os quantis de ordem 0, 99 ($x_{0,01}$) e $0,999$ ($x_{0,001}$) estão se comportando (probabilidade de cobertura, amplitudes médias e limites inferior e superior médios) com o aumento do tamanho amostral, utilizando k_1 e k_2.

Começando pela distribuição Weibull, devemos ressaltar que os quantis de ordem $0,99$ e $0,999$, para os parâmetros $\beta = 0,3$ e $\alpha = 1$, correspondem a 162,4871 e 627,7545, respectivamente. Os resultados dos intervalos de confiança, para tais quantis, encontram-se nas Tabelas 3.7, 3.8, 3.9 e 3.10. Contudo, achamos que gráficos possam vir a ser mais informativos do que tabelas. Desse modo, os mesmos resultados podem ser encontrados nas Figuras 3.8, 3.9 e 3.10, das quais podemos observar que as melhores probabilidades de cobertura dos intervalos vêm sendo obtidas utilizando k_1 para os quantis de ordem $0,99$. Todavia, para os quantis de ordem $0,999$, as probabilidades de cobertura diminuem bastante ao se utilizar k_1. Já o uso de k_2, para os cálculos dos intervalos, mostra resultados mais parecidos entre os quantis de ordem $0,99$ e $0,999$. Ainda, as amplitudes médias

dos intervalos, com exceção, de $n = 1000$ para os quantis de ordem $0,999$, foram sempre menores ao utilizarmos k_2. Por fim, para os quantis de ordem $0,999$, devemos observar que os intervalos de confiança contêm em média o verdadeiro quantil de ordem $0,999$ somente empregando em k_2 e com tamanho amostral $n = 2000$. Para os quantis de ordem $0,99$ observamos que, tanto utilizando k_1 como k_2, os intervalos de confiança apresentam um comportamento parecido em todos os métodos. Em decorrência disso, para os dados oriundos de distribuição Weibull, acreditamos que intervalos utilizando o valor de k que minimiza o erro quadrático médio assintótico do estimador de Hill, k_2, fornecem melhores resultados.

Da mesma forma como foi feito com as amostras de distribuição Weibull, fizemos com as amostras de distribuição Fréchet. Não é difícil obter que os quantis de ordem $0,99$ e $0,999$ de uma distribuição Fréchet de parâmetro $\alpha = 1$ correspondem a 99,49916 e 999,49992, respectivamente. Os resultados referentes aos intervalos de confiança para estes quantis encontram-se nas Tabelas 3.11, 3.12, 3.13 e 3.14. Resultados análogos encontram-se nas Figuras 3.11, 3.12 e 3.13. Como já comentamos, achamos os gráficos mais informativos do que as tabelas, por isso vamos analisar os intervalos por meio deles. Na Figura 3.11 podemos observar que, utilizando k_1, temos melhores probabilidades de cobertura para ambos os quantis, $x_{0,01}$ e $x_{0,001}$, do que utilizando k_2. Provavelmente isso ocorre pelo fato de tais intervalos possuírem, em média, amplitudes maiores, como exibido na Figura 3.12. Na Figura 3.13 podemos observar que, para ambos os quantis, $x_{0,01}$ e $x_{0,001}$, todos os intervalos calculados utilizando tanto k_1 quanto k_2 contêm, em média, o verdadeiro valor do quantil. Em decorrência destes resultados, acreditamos que intervalos obtidos utilizando $k_1 = 1,5(\log n)^2$, para os dados oriundos de distribuição Fréchet, fornecem os melhores resultados.

Com base nas simulações, temos indícios de que, se os dados possuírem distribuição com cauda de variação regular (ver Definição 2), a melhor escolha será baseada em $k_1 = 1,5(\log n)^2$. Caso contrário, para os três métodos utilizados aqui (aproximação pela normal, razão de verossimilhança e *data tilting*), k_2 auxiliará na obtenção de intervalos mais robustos à quebra da suposição (2.2), para dados com distribuição subexponencial (ver Definição 1).

3.3 Comparação dos métodos

Para comparar os métodos utilizados nesta dissertação, faremos uso dos resultados exibidos nas Seções 3.1 e 3.2 deste capítulo. Com relação ao método da gama generalizada, não obtivemos os intervalos para as amostras da distribuição Fréchet, por problemas no processo de estimação dos

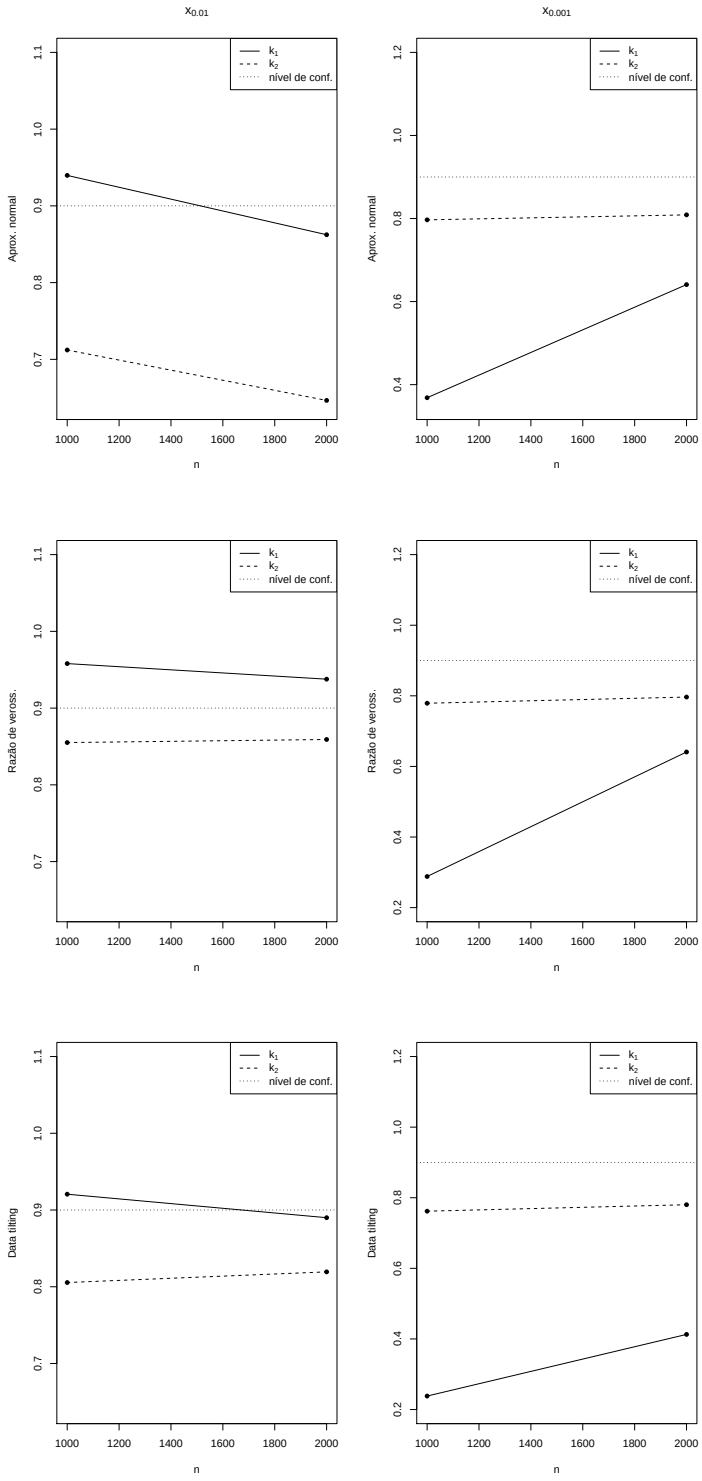

Figura 3.8 *Probabilidades de cobertura dos intervalos com 90% de confiança para os quantis de ordem* $0,99$ *(*$x_{0,01}$*) – lado esquerdo – e* $0,999$ *(*$x_{0,001}$*) – lado direito – da distribuição Weibull, comparando as escolhas* k_1 *e* k_2*, para os métodos da aproximação pela normal, razão de verossimilhanças e data tilting.*

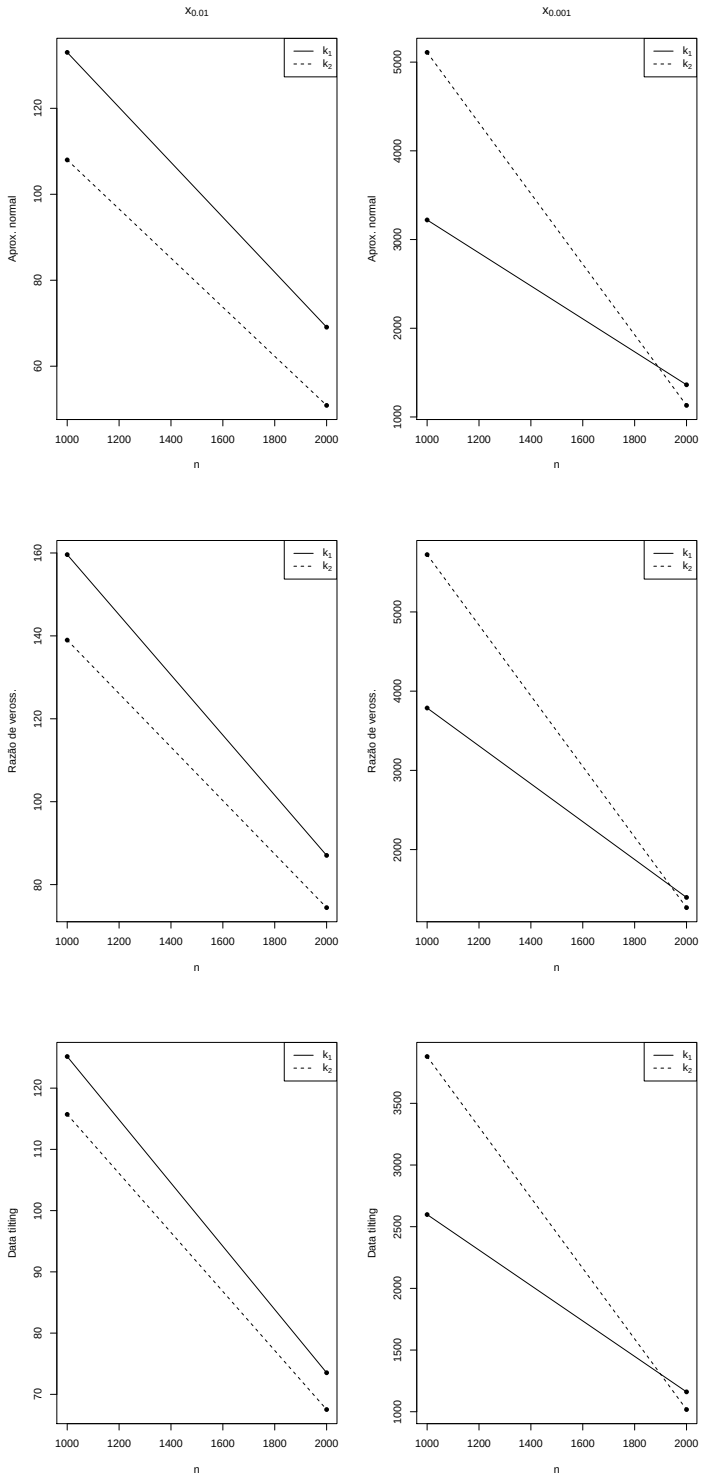

Figura 3.9 *Amplitudes médias dos intervalos com 90% de confiança para os quantis de ordem* $0,99$ *(*$x_{0,01}$*) – lado esquerdo – e* $0,999$ *(*$x_{0,001}$*) – lado direito – da distribuição Weibull, comparando as escolhas* k_1 *e* k_2*, para os métodos da aproximação pela normal, razão de verossimilhanças e data tilting.*

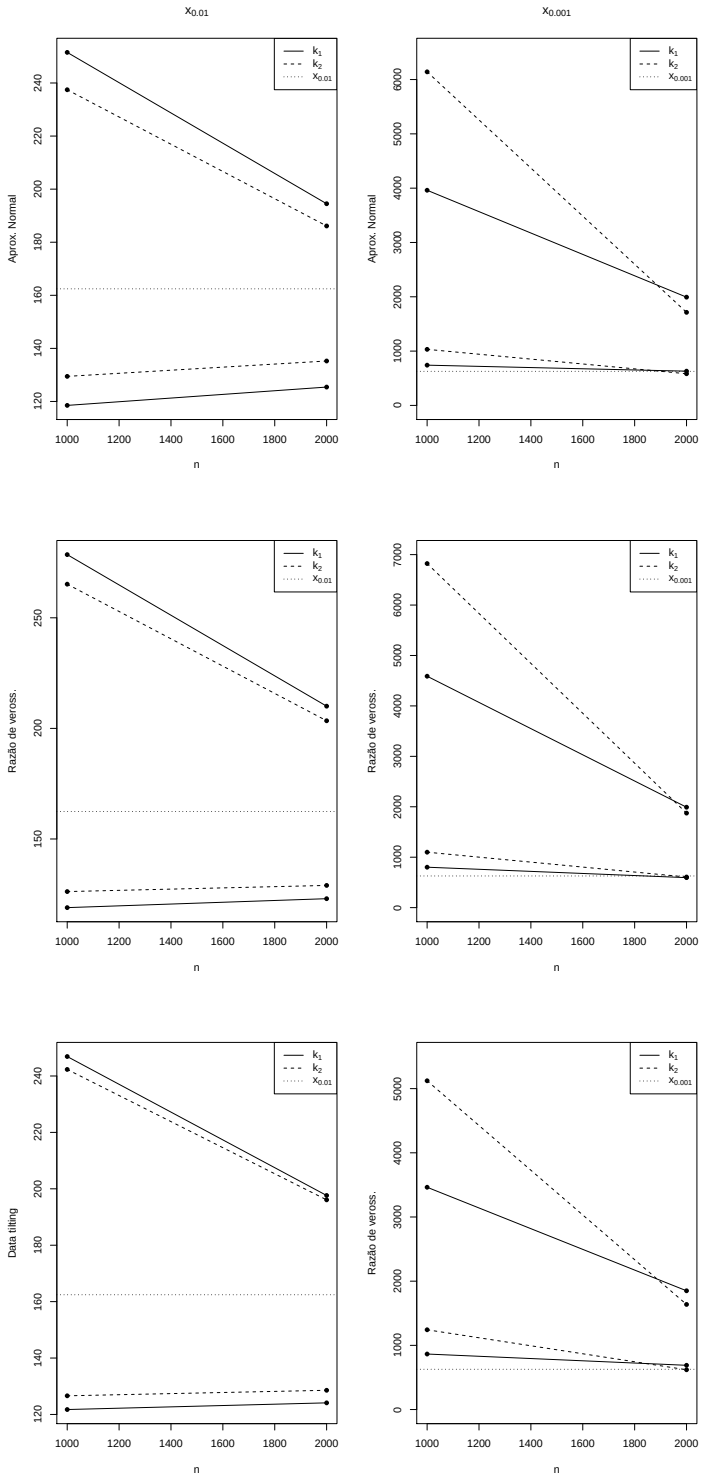

Figura 3.10 *Limites inferiores e superiores médios dos intervalos com 90% de confiança para os quantis de ordem* $0,99$ *(*$x_{0,01}$*) – lado esquerdo – e* $0,999$ *(*$x_{0,001}$*) – lado direito – da distribuição Weibull, comparando as escolhas* k_1 *e* k_2*, para os métodos da aproximação pela normal, razão de verossimilhanças e data tilting.*

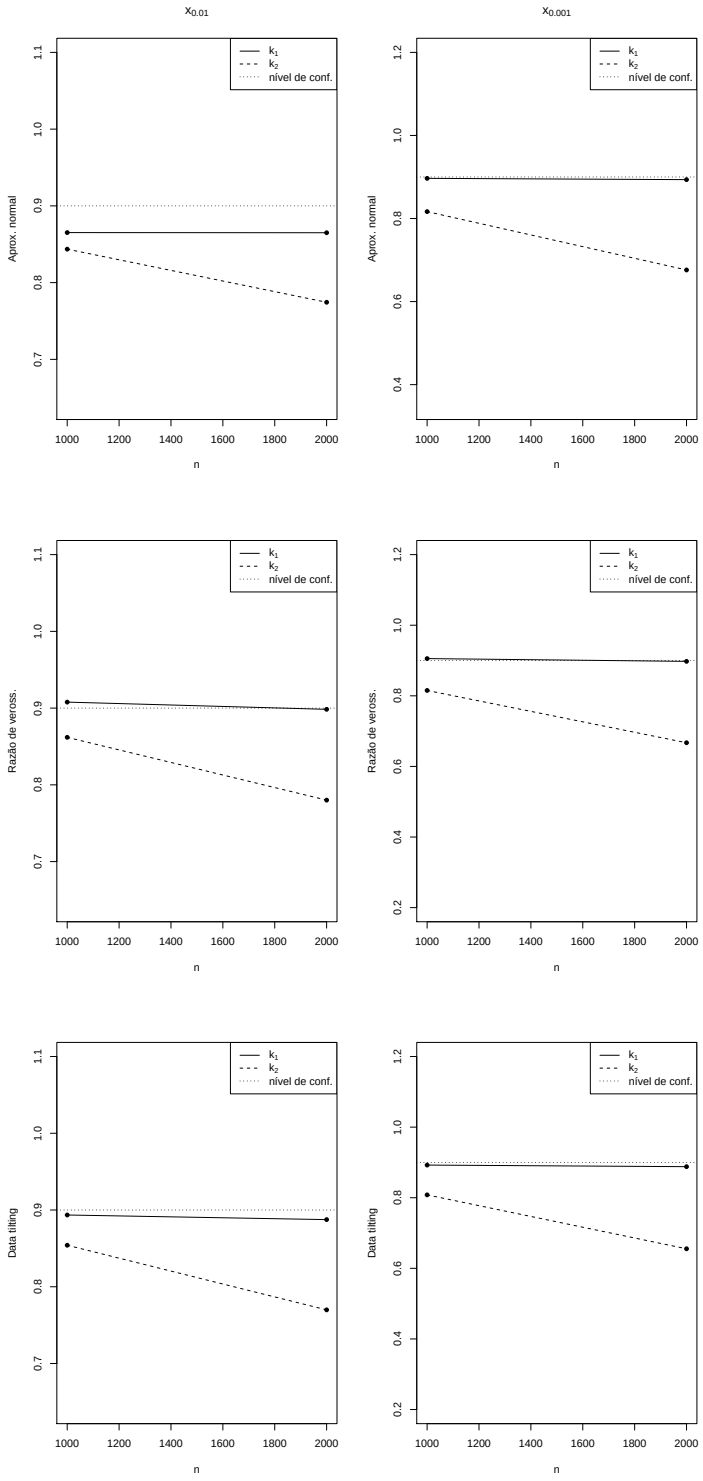

Figura 3.11 *Probabilidades de cobertura dos intervalos com 90% de confiança para os quantis de ordem* $0,99$ *(*$x_{0,01}$*) – lado esquerdo – e* $0,999$ *(*$x_{0,001}$*) – lado direito – da distribuição Fréchet, comparando as escolhas* k_1 *e* k_2*, para os métodos da aproximação pela normal, razão de verossimilhanças e data tilting.*

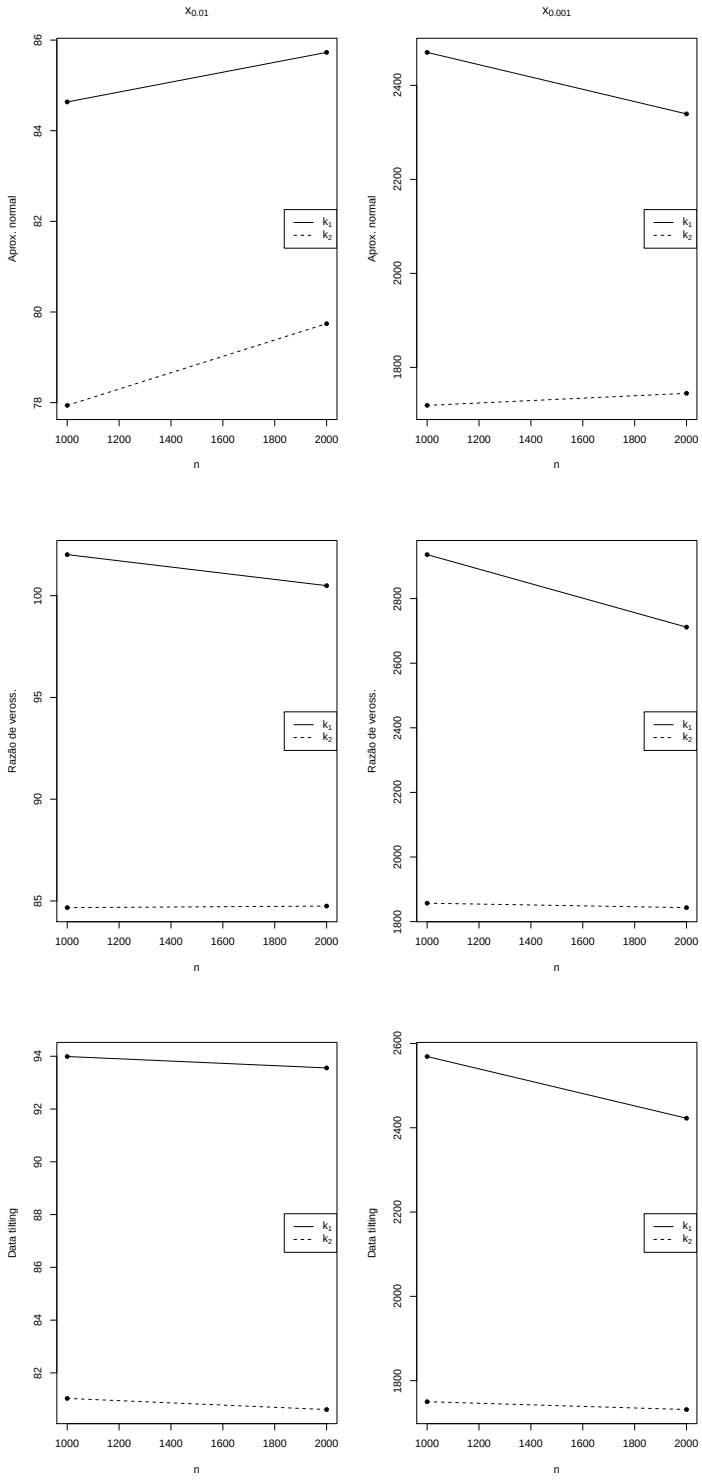

Figura 3.12 *Amplitudes médias dos intervalos com 90% de confiança para os quantis de ordem* $0,99$ *(*$x_{0,01}$*) – lado esquerdo – e* $0,999$ *(*$x_{0,001}$*) – lado direito – da distribuição Fréchet, comparando as escolhas* k_1 *e* k_2*, para os métodos da aproximação pela normal, razão de verossimilhanças e data tilting.*

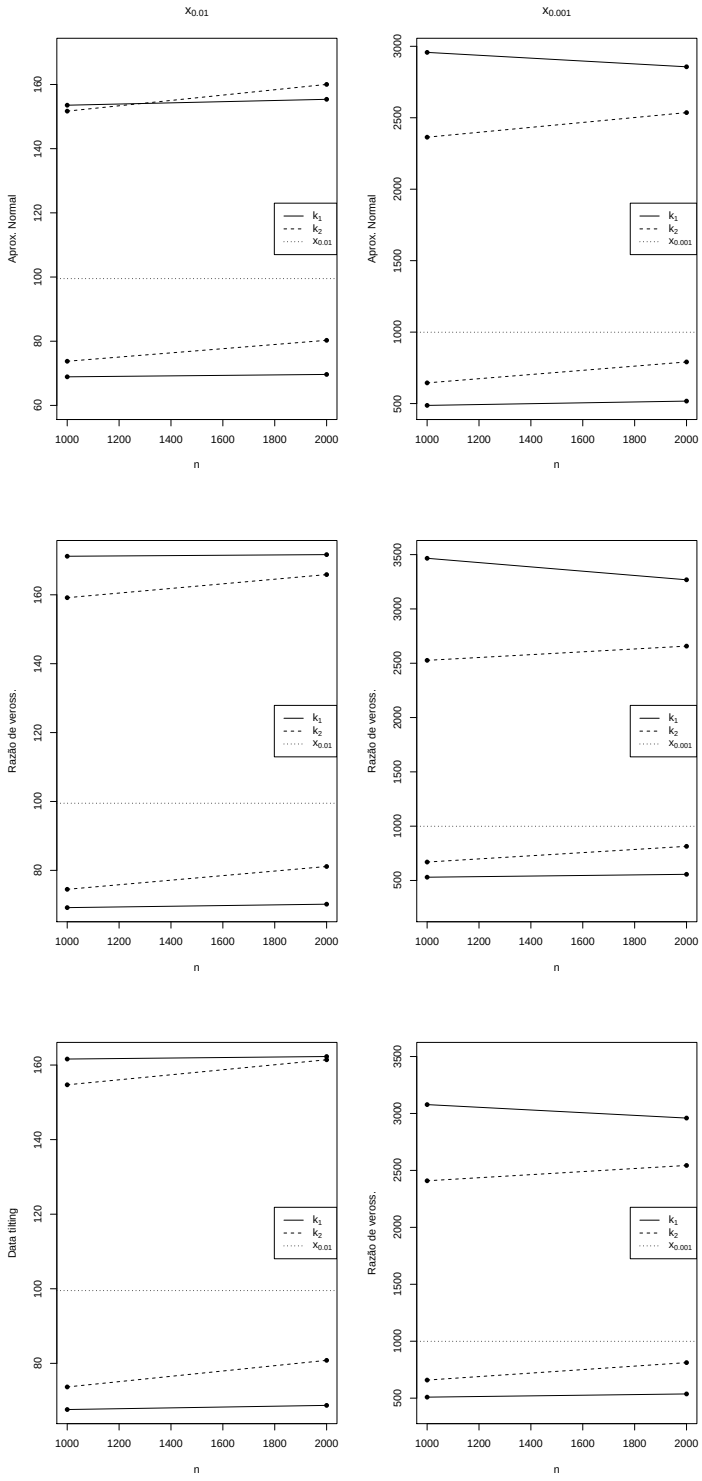

Figura 3.13 *Limites inferiores e superiores médios dos intervalos com 90% de confiança para os quantis de ordem* $0,99$ *(*$x_{0,01}$*) – lado esquerdo – e* $0,999$ *(*$x_{0,001}$*) – lado direito – da distribuição Fréchet, comparando as escolhas* k_1 *e* k_2*, para os métodos da aproximação pela normal, razão de verossimilhanças e data tilting.*

		Aprox. normal		Razão de Veross.		Data Tilting	
n	k	Cobert.	Amplitude	Cobert.	Amplitude	Cobert.	Amplitude
1000	k_1	0,9397	133,0126	0,9580	159,6032	0,9206	125,1596
	k_2	0,7122	107,9937	0,8551	138,9842	0,8055	115,7127
2000	k_1	0,8624	69,0804	0,9377	87,0429	0,8900	73,5287
	k_2	0,6465	50,8795	0,8591	74,4457	0,8194	67,5201

Tabela 3.7 *Probabilidades de coberturas e amplitudes médias dos intervalos com 90% de confiança para o quantil de ordem* $0,99$ $(x_{0,01} = 162,4871)$*, para 10000 amostras de tamanhos 1000 e 2000 da distribuição Weibull, utilizando* k_1 *e* k_2*.*

		Aprox. normal		Razão de Veross.		Data Tilting	
n	k	Cobert.	Amplitude	Cobert.	Amplitude	Cobert.	Amplitude
1000	k_1	0,3685	3220,7990	0,2882	3787,2470	0,2382	2598,2870
	k_2	0,7968	5109,6770	0,7790	5723,3810	0,7619	3879,4220
2000	k_1	0,6409	1362,7270	0,6409	1396,3170	0,4128	1161,0070
	k_2	0,8088	1130,3380	0,7965	1267,4400	0,7802	1017,7510

Tabela 3.8 *Probabilidades de coberturas e amplitudes médias dos intervalos com 90% de confiança para o quantil de ordem* $0,999$ $(x_{0,001} = 627,7545)$*, para 10000 amostras de tamanhos 1000 e 2000 da distribuição Weibull, utilizando* k_1 *e* k_2*.*

parâmetros (tanto nas amostras de tamanho 1000, quanto nas amostras de tamanho 2000, em mais de 98% delas, o processo iterativo de estimação dos parâmetros não convergiu). Com relação aos outros três métodos, utilizamos o limiar baseado em $k_1 = 1,5(\log n)^2$ para as amostras da distribuição Fréchet, e o valor de k que minimiza o erro quadrático médio assintótico, k_2, para as amostras da distribuição Weibull.

Podemos observar na Figura 3.14 que, para as amostras geradas da distribuição Weibull, o método

		Aprox. normal		Razão de Veross.		Data Tilting	
n	k	Lim. Inf.	Lim. Sup.	Lim. Inf.	Lim. Sup.	Lim. Inf.	Lim. Sup.
1000	k_1	118,5139	251,5265	118,9624	278,5657	121,7400	246,8996
	k_2	129,4529	237,4466	126,2219	265,2062	126,6089	242,3216
2000	k_1	125,4205	194,5009	123,0069	210,0498	124,1096	197,6383
	k_2	135,2469	186,1264	129,0034	203,4490	128,5663	196,0865

Tabela 3.9 *Limites inferior e superior médios dos intervalos com 90% de confiança para o quantil de ordem* $0,99$ $(x_{0,01} = 162,4871)$*, para 10000 amostras de tamanhos 1000 e 2000 da distribuição Weibull, utilizando* k_1 *e* k_2*.*

		Aprox. normal		Razão de Veross.		Data Tilting	
n	k	Lim. Inf.	Lim. Sup.	Lim. Inf.	Lim. Sup.	Lim. Inf.	Lim. Sup.
1000	k_1	740,7940	3961,5930	801,5584	4588,8055	864,8866	3463,1740
	k_2	1032,1790	6141,8560	1099,4610	6822,8420	1242,7900	5122,2120
2000	k_1	630,3083	1993,0351	594,8652	1991,1823	689,2521	1850,2596
	k_2	583,0539	1713,3918	607,0158	1874,4562	619,8161	1637,5667

Tabela 3.10 *Limites inferior e superior médios dos intervalos com 90% de confiança para o quantil de ordem* $0,999$ ($x_{0,001} = 627,7545$)*, para 10000 amostras de tamanhos 1000 e 2000 da distribuição Weibull, utilizando* k_1 *e* k_2*.*

		Aprox. normal		Razão de Veross.		Data Tilting	
n	k	Cobert.	Amplitude	Cobert.	Amplitude	Cobert.	Amplitude
1000	k_1	0,8652	84,6344	0,9078	102,0175	0,8935	93,9896
	k_2	0,8436	77,9392	0,8619	84,6711	0,8541	81,0334
2000	k_1	0,8650	85,7281	0,8984	100,4927	0,8875	93,5579
	k_2	0,7744	79,7428	0,7801	84,7476	0,8875	93,5579

Tabela 3.11 *Probabilidades de coberturas e amplitudes médias dos intervalos com 90% de confiança para o quantil de ordem* $0,99$ ($x_{0,01} = 99,49916$)*, para 10000 amostras de tamanhos 1000 e 2000 da distribuição Fréchet, utilizando* k_1 *e* k_2*.*

da gama generalizada mostrou-se superior aos demais, pois possui as melhores probabilidades de cobertura, as menores amplitudes médias (com excessão no quantil de ordem $0,99$ quando $n = 2000$, onde o método da aproximação pela normal apresenta a menor amplitude média) e é o único método onde os intervalos contêm, em média, o verdadeiro valor do quantil $x_{0,001}$. Assim, como já esperávamos, o método da gama generalizada, que contém a distribuição Weibull como caso particular, fornece melhores resultados no cálculo dos intervalos de confiança.

		Aprox. normal		Razão de Veross.		Data Tilting	
n	k	Cobert.	Amplitude	Cobert.	Amplitude	Cobert.	Amplitude
1000	k_1	0,8967	2470,187	0,9055	2936,181	0,8925	2568,970
	k_2	0,8166	1719,110	0,8152	1856,986	0,8081	1850,070
2000	k_1	0,8936	2339,305	0,8976	2711,623	0,8880	2422,585
	k_2	0,6761	1744,814	0,6671	1843,241	0,6554	1731,461

Tabela 3.12 *Probabilidades de coberturas e amplitudes médias dos intervalos com 90% de confiança para o quantil de ordem* $0,999$ ($x_{0,001} = 999,49992$)*, para amostras de tamanhos 1000 e 2000 da distribuição Fréchet, utilizando* k_1 *e* k_2*.*

		Aprox. normal		Razão de Veross.		Data Tilting	
n	k	Lim. Inf.	Lim. Sup.	Lim. Inf.	Lim. Sup.	Lim. Inf.	Lim. Sup.
1000	k_1	68,9062	153,5406	69,1778	171,1954	67,6205	161,6101
	k_2	73,7563	151,6956	74,4944	159,1655	73,6674	154,7008
2000	k_1	69,6506	155,3787	70,1761	170,6687	68,7246	161,2825
	k_2	80,2803	160,0231	81,1214	165,8690	80,8029	161,4146

Tabela 3.13 *Limites inferior e superior médios dos intervalos com 90% de confiança para o quantil de ordem* $0,99$ ($x_{0,01} = 99,49916$)*, para amostras de tamanhos 1000 e 2000 da distribuição Fréchet, utilizando* k_1 *e* k_2*.*

		Aprox. normal		Razão de Veross.		Data Tilting	
n	k	Lim. Inf.	Lim. Sup.	Lim. Inf.	Lim. Sup.	Lim. Inf.	Lim. Sup.
1000	k_1	487,1492	2957,3364	530,0434	3466,2243	508,8264	3077,7966
	k_2	644,7126	2363,8229	669,6805	2526,6664	658,7503	2408,8207
2000	k_1	517,2102	2856,5155	556,7384	3268,3610	536,9880	2959,5730
	k_2	791,1886	2536,0030	814,3447	2657,5859	812,0182	2543,4795

Tabela 3.14 *Limites inferior e superior médios dos intervalos com 90% de confiança para o quantil de ordem* $0,999$ ($x_{0,001} = 999,49992$)*, para amostras de tamanhos 1000 e 2000 da distribuição Fréchet, utilizando* k_1 *e* k_2*.*

Para o caso das amostras geradas com distribuição Fréchet, podemos verificar na Figura 3.15 que, utilizando $k_1 = 1,5(\log n)^2$, o método da razão de verossimilhanças apresenta as probabilidades de cobertura mais próximas do nível de confiança (0,90). Provavelmente isso ocorra pelo fato deste método possuir amplitudes médias maiores do que os demais métodos. Nesta mesma figura podemos observar também que, todos os métodos contêm, em média, o verdadeiro quantil de ordem $0,99$ em seus intervalos de confiança. Desse modo, utilizando k_1, o método da razão de verossimilhanças fornece os melhores resultados no cálculo dos intervalos de confiança, para a distribuição Fréchet.

Em decorrência dos resultados obtidos nesta seção, acreditamos que seja de suma importância o uso de algum tipo de análise exploratória dos dados, com o intuito de se ter alguma idéia do tipo de distribuição da amostra, antes de se utilizar as técnicas de estimação intervalar. No Capítulo 4 encontram-se algumas técnicas gráficas que auxiliam na escolha entre distribuições subexponenciais (sem que as mesmas pertençam à classe de distribuições com cauda de variação regular) e distribuições com caudas de variação regular. Tais técnicas serão utilizadas no Capítulo 4, para um conjunto de dados reais.

	$n = 1000$		$n = 2000$	
	Cobertura	Amplitude	Cobertura	Amplitude
Aprox. normal	0,7122	107,9937	0,6465	50,8795
Razão de veross.	0,8551	138,9842	0,8591	74,4457
Data tilting	0,8055	115,7127	0,8194	67,5201
Gama generalizada	0,9108	90,2123	0,9049	56,4704

Tabela 3.15 *Probabilidades de coberturas e amplitudes médias dos intervalos com 90% de confiança para o quantil de ordem* $0,99$ ($x_{0,01} = 162,4871$)*, para 10000 amostras de tamanhos 1000 e 2000 da distribuição Weibull, utilizando* k_2 *nos métodos da aproximação pela normal, razão de verossimilhanças e data tilting.*

	$n = 1000$		$n = 2000$	
	Cobertura	Amplitude	Cobertura	Amplitude
Aprox. normal	0,7968	5109,6770	0,8088	1130,3380
Razão de veross.	0,7790	5723,3810	0,7965	1267,4400
Data tilting	0,7619	3879,4220	0,7802	1017,7510
Gama generalizada	0,9153	411,5739	0,9096	342,4411

Tabela 3.16 *Probabilidades de coberturas e amplitudes médias dos intervalos com 90% de confiança para o quantil de ordem* $0,999$ ($x_{0,001} = 627,7545$)*, para 10000 amostras de tamanhos 1000 e 2000 da distribuição Weibull, utilizando* k_2 *nos métodos da aproximação pela normal, razão de verossimilhanças e data tilting.*

	$n = 1000$		$n = 2000$	
	Cobertura	Amplitude	Cobertura	Amplitude
Aprox. normal	0,8652	84,6344	0,8650	85,7281
Razão de veross.	0,9078	102,0175	0,8984	100,4927
Data tilting	0,8935	93,9896	0,8875	93,5579

Tabela 3.17 *Probabilidades de coberturas e amplitudes médias dos intervalos com 90% de confiança para o quantil de ordem* $0,99$ ($x_{0,01} = 99,49916$)*, para 10000 amostras de tamanhos 1000 e 2000 da distribuição Fréchet, utilizando* k_1 *nos métodos da aproximação pela normal, razão de verossimilhanças e data tilting.*

	$n = 1000$		$n = 2000$	
	Cobertura	Amplitude	Cobertura	Amplitude
Aprox. normal	0,8967	2470,187	0,8936	2339,305
Razão de veross.	0,9055	2936,181	0,8976	2711,623
Data tilting	0,8925	2568,970	0,8880	2422,585

Tabela 3.18 *Probabilidades de coberturas e amplitudes médias dos intervalos com 90% de confiança para o quantil de ordem* $0,999$ ($x_{0,001} = 999,49992$)*, para 10000 amostras de tamanhos 1000 e 2000 da distribuição Fréchet, utilizando* k_1 *nos métodos da aproximação pela normal, razão de verossimilhanças e data tilting.*

	$n = 1000$		$n = 2000$	
	Lim. Inf.	Lim. Sup.	Lim. Inf.	Lim. Sup.
Aprox. normal	129,4529	237,4466	135,4269	186,1264
Razão de veross.	126,2219	265,2062	129,0034	203,4490
Data tilting	126,6089	242,3216	128,5663	196,0865
Gama generalizada	117,3985	208,1508	131,0232	194,4175

Tabela 3.19 *Limites inferior e superior médios dos intervalos com 90% de confiança para o quantil de ordem* $0,99$ ($x_{0,01} = 162,4871$)*, para 10000 amostras de tamanhos 1000 e 2000 da distribuição Weibull, utilizando* k_2 *nos métodos da aproximação pela normal, razão de verossimilhanças e data tilting.*

	$n = 1000$		$n = 2000$	
	Lim. Inf.	Lim. Sup.	Lim. Inf.	Lim. Sup.
Aprox. normal	1032,1790	6141,8560	583,0539	1713,3918
Razão de veross.	1099,4610	6822,8420	607,0158	1874,4562
Data tilting	1242,7900	5122,2120	619,8161	1637,5667
Gama generalizada	389,8645	881,4383	459,4826	801,9237

Tabela 3.20 *Limites inferior e superior médios dos intervalos com 90% de confiança para o quantil de ordem* $0,999$ ($x_{0,001} = 627,7545$)*, para 10000 amostras de tamanhos 1000 e 2000 da distribuição Weibull, utilizando* k_2 *nos métodos da aproximação pela normal, razão de verossimilhanças e data tilting.*

	$n = 1000$		$n = 2000$	
	Lim. Inf.	Lim. Sup.	Lim. Inf.	Lim. Sup.
Aprox. normal	68,9062	162,5406	69,6506	155,3787
Razão de veross.	69,1778	171,1954	70,1761	170,6687
Data tilting	67,6205	161,6101	68,7246	161,2825

Tabela 3.21 *Limites inferior e superior médios dos intervalos com 90% de confiança para o quantil de ordem* $0,99$ ($x_{0,01} = 99,49916$)*, para 10000 amostras de tamanhos 1000 e 2000 da distribuição Fréchet, utilizando* k_1 *nos métodos da aproximação pela normal, razão de verossimilhanças e data tilting.*

	$n = 1000$		$n = 2000$	
	Lim. Inf.	Lim. Sup.	Lim. Inf.	Lim. Sup.
Aprox. normal	487,1492	2957,3364	517,2102	2856,5155
Razão de veross.	530,0434	3466,2243	556,7384	3268,3610
Data tilting	508,8264	3077,7966	536,9880	2959,5730

Tabela 3.22 *Limites inferior e superior médios dos intervalos com 90% de confiança para o quantil de ordem* $0,999$ ($x_{0,001} = 999,49992$)*, para 10000 amostras de tamanhos 1000 e 2000 da distribuição Fréchet, utilizando* k_1 *nos métodos da aproximação pela normal, razão de verossimilhanças e data tilting.*

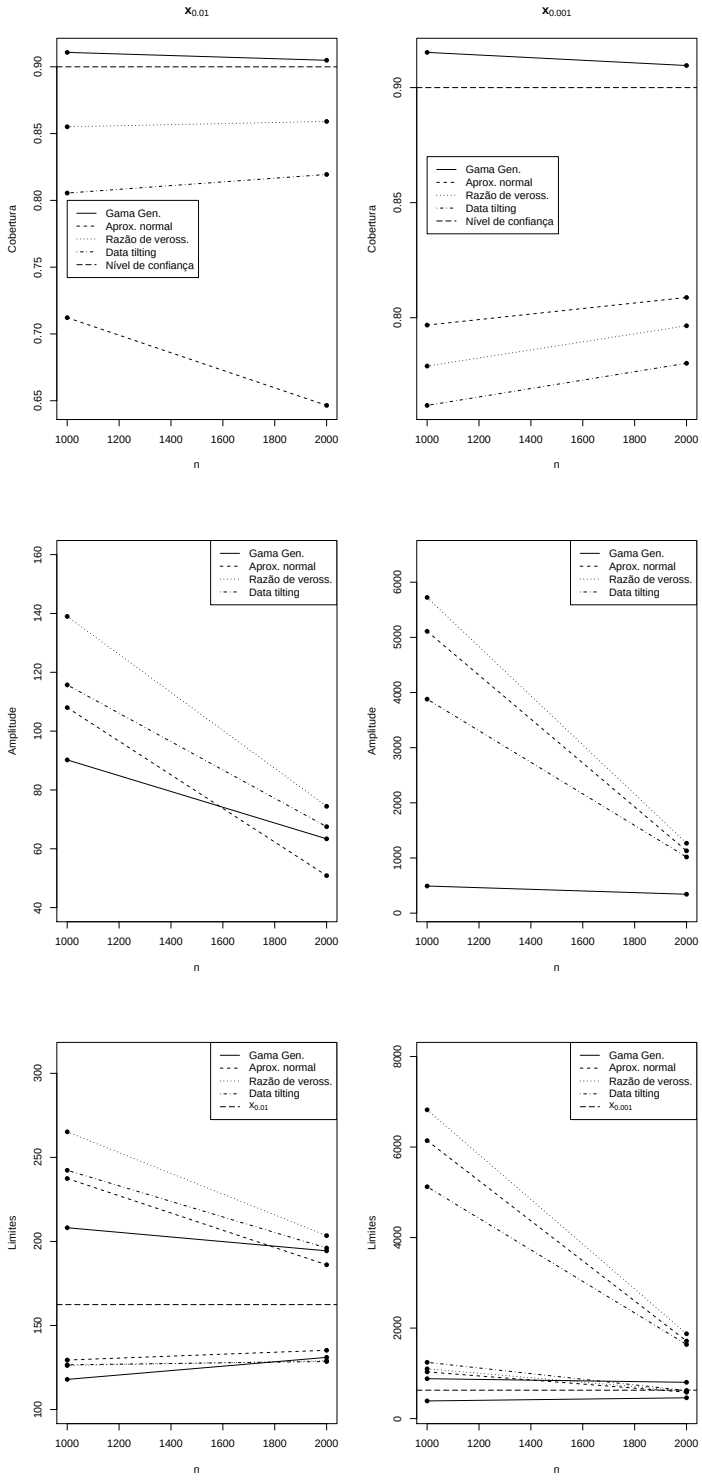

Figura 3.14 *Probabilidades de cobertura dos intervalos com 90% de confiança para os quantis de ordem* $0,99$ *(*$x_{0,01}$*) e* $0,999$ *(*$x_{0,001}$*) da distribuição Weibull, comparando os métodos da aproximação pela normal, razão de verossimilhanças, data tilting e gama generalizada.*

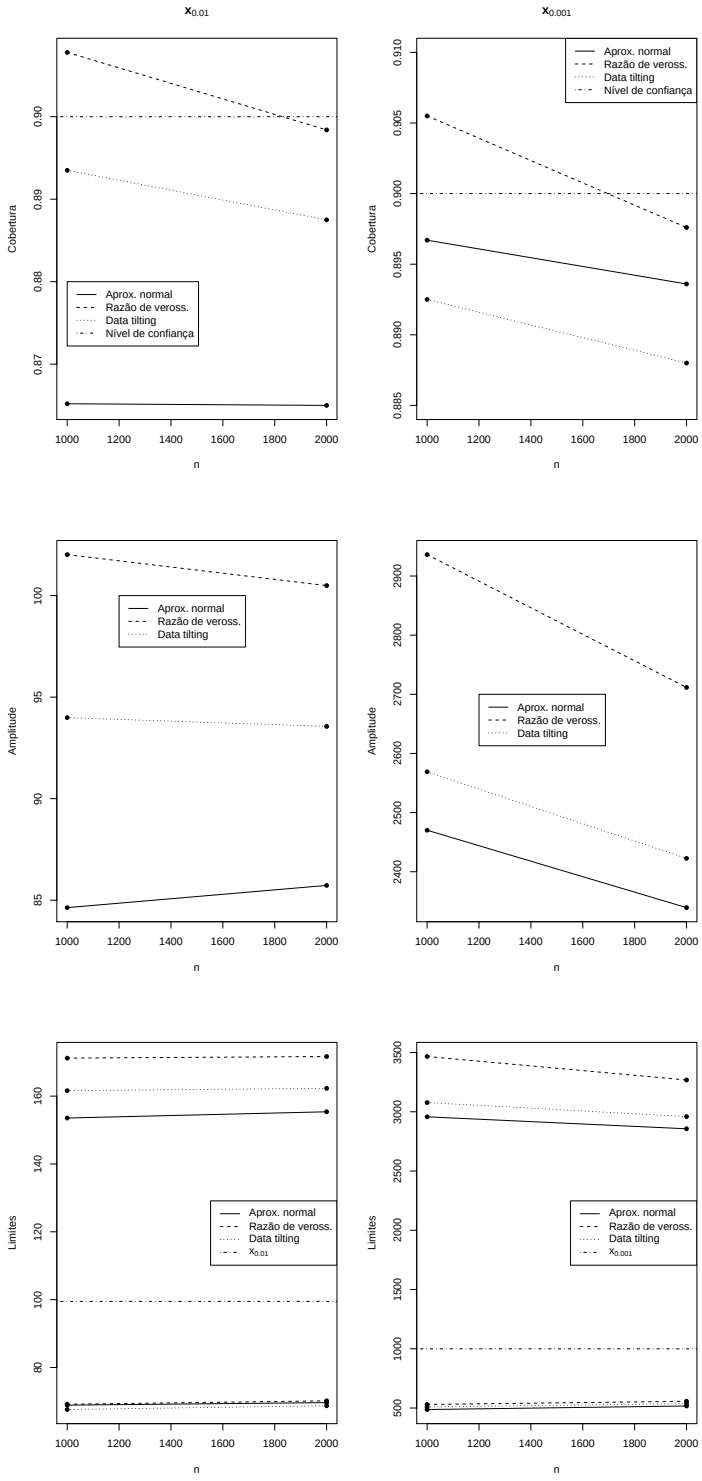

Figura 3.15 *Probabilidades de cobertura dos intervalos com 90% de confiança para os quantis de ordem* $0,99$ *(*$x_{0,01}$*) e* $0,999$ *(*$x_{0,001}$*) da distribuição Fréchet, comparando os métodos da aproximação pela normal, razão de verossimilhanças e data tilting.*

Capítulo 4

Aplicação

Como comentamos na Seção 1.1, grandes sinistros podem colocar em risco a solvência de um portfolio ou, até mesmo, de uma parte substancial da companhia de seguros, daí a necessidade da estimação de altos quantis para valores de sinistros. Em decorrência do fato das estimativas pontuais nem sempre fornecerem resultados muito próximos do verdadeiro valor do quantil de interesse, nos interessamos em obter intervalos de confiança para tais quantis.

Neste capítulo analisaremos um conjunto de 1758 dados reais, os mesmos utilizados em Melo (2006), referentes ao pagamento de indenizações em reais (por sinistros) de seguros de incêndio, de um determinado grupo de seguradoras no Brasil, no ano de 2003. Tais dados, que podem ser observados na Figura 4.1, serão utilizados para obtermos intervalos com 90% de confiança para os quantis de ordem $0,99$ e $0,999$, e verificarmos o comportamento de tais intervalos, sob os métodos utilizados neste trabalho (ver Capítulo 2).

Inicialmente, exibimos na Tabela 4.1 algumas estatísticas descritivas deste conjunto de dados, assim como na Figura 4.2 o histograma, onde podemos observar, principalmente, que os dados apresentam uma forte assimetria.

Mínimo	1,00
1^o Quartil	500,00
Mediana	2332,50
3^o Quartil	12020,10
Máximo	3041790,00
Média	25199,36
Coef. Assimetria	15,93
Coef. Curtose	338,08

Tabela 4.1 *Estatísticas descritivas referentes ao pagamento de indenizações de seguros de incêndio no Brasil em 2003*

Com o intuito de facilitar na identificação da classe de distribuição dos dados, assim como pro-

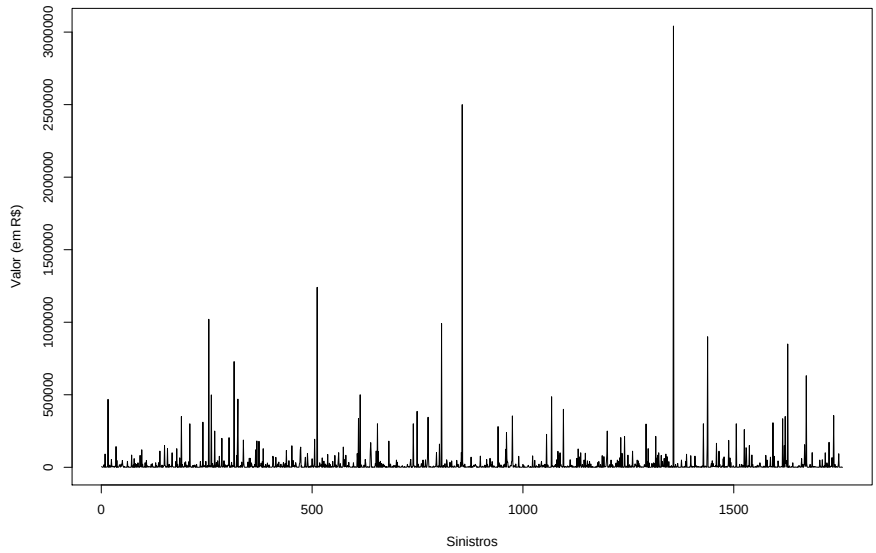

Figura 4.1 *Dados referentes ao pagamento de indenizações de seguros de incêndio no Brasil em 2003.*

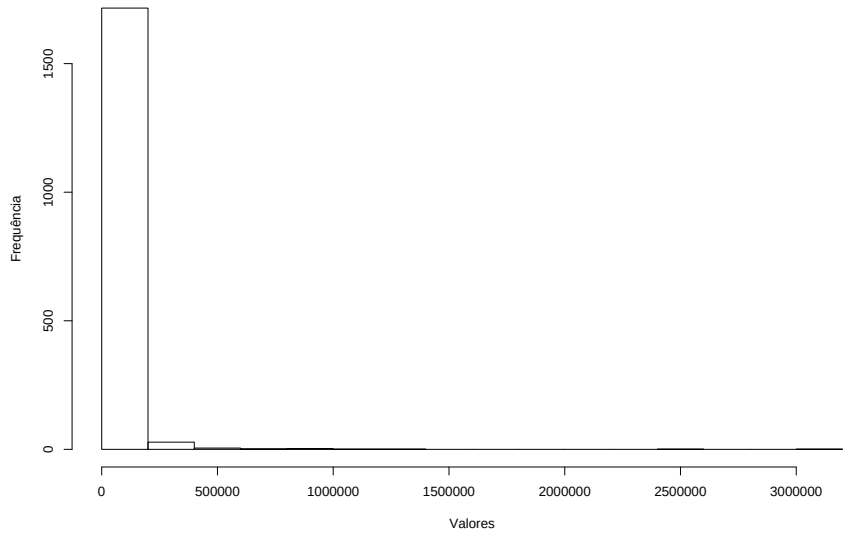

Figura 4.2 *Histograma dos dados referentes ao pagamento de indenizações de seguros de incêndio no Brasil em 2003.*

posto por Adlouni *et al.* (2008), utilizaremos algumas técnicas gráficas, as quais serão comentadas em seguida.

Gráfico log–log. Uma das suposições para as distribuições com cauda de variação regular que estamos fazendo aqui é que

$$\mathbb{P}(X > x) = cx^{-\gamma}, x > T$$

ou seja, para valores grandes de x teremos

$$\log \mathbb{P}(X > x) = \log c - \gamma \log x$$

Desse modo, para distribuições com cauda de variação regular, esperamos que para alguns valores elevados de x, o gráfico dos pontos $(\log x, \log \mathbb{P}(X > x))$ se comporte de forma linear e decrescente, ao contrário das distribuições subexponenciais que não possuem caudas de variação regular. Um exemplo da diferença entre estes dois tipos de distribuições pode ser visualizado na Figura 4.3.

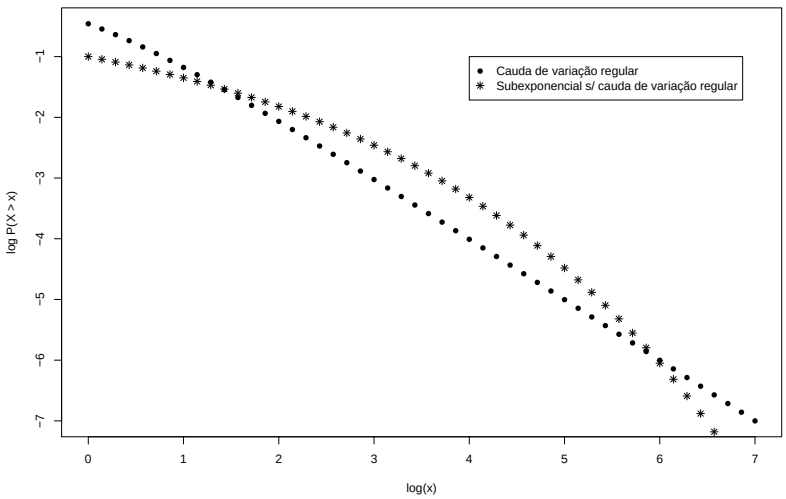

Figura 4.3 *Comparação entre gráficos log–log para distribuições com cauda de variação regular e distribuições subexponenciais que não possuem cauda de variação regular.*

Gráfico de Hill. Esta ferramente gráfica visa estimar $\hat{\gamma}_n$, em (2.9), para vários valores de k. O que se espera, para dados oriundos de uma distribuição com cauda de variação regular, é que com o aumento de k, as estimativas de Hill, $\hat{\gamma}_n$, tenham valores próximos, ou seja, se comportem de maneira estável.

Gráfico da razão entre máximos e somas. Esta ferramenta consiste basicamente na estatística $R_n(p) = \frac{\max\{X_1^p,\ldots,X_n^p\}}{\sum_{i=1}^n X_i^p}$, a qual convergirá para zero quando $n \to \infty$, para séries estacionárias, se e

somente se $\mathbb{E}(X^p) < \infty$. Tem-se ainda, para dados de distribuição com cauda de variação regular, que se p for maior do que o índice da cauda (*tail index*), γ, $R_n(p)$ não converge para zero. Dessa forma, o que se faz na prática é construir gráficos entre $(i, R_i(p))$, $i = 1, 2, \ldots, n$, para alguns valores de p. Se em todos os gráficos, observarmos uma aparente convergência de $R_n(p)$, teremos indícios de que os dados não possuem distribuição com cauda de variação regular.

Antes de calcularmos os intervalos de confiança, tentaremos obter indícios de qual tipo de distribuição que o conjunto de dados deve apresentar, utilizando as ferramentas gráficas supracitadas, que são úteis para tal finalidade. A finalidade de tais gráficos é auxiliar na diferenciação entre distribuições com cauda de variação regular (ver Definição 2) e distribuições subexponenciais (ver Definição 1) sem que estas pertençam à classe de distribuições com cauda de variação lenta. Desse modo, podemos observar na Figura 4.4, nas últimas estatísticas de ordem, que tal gráfico aparenta um decaimento linear, dando-nos indícios de que tal conjunto de dados pertence à classe de distribuições com cauda de variação regular. Tal indício é reforçado ao observarmos a Figura 4.5, onde podemos ver que as estimativas de Hill apresentam uma aparente estabilização quando o valor de k aumenta, o que é esperado para dados provindos de distribuições com cauda de variação regular. Na Figura 4.6 podemos perceber que, em $p = 1$, temos que a razão entre a soma e o máximo dos dados tende a convergir para zero, à medida que n aumenta, dando-nos também fortes indícios de que os dados referentes aos pagamentos de indenizações de seguros de incêndios no Brasil pertençam à classe de distribuições com cauda de variação regular, sendo talvez, mais adequado o uso dos métodos aproximação pela normal, razão de verossimilhanças e *data tilting*, do que o método da gama generalizada.

Também podemos tentar verificar se os dados apresentam distribuição Gama Generalizada. Para tanto, estimamos os parâmetros da distribuição Gama Generalizada, de acordo com o Capítulo 2.4, o que resultou em $\hat{\beta} = 0,1901$, $\hat{\zeta} = 6,1636$ e $\hat{\alpha} = 0.2890$. Na Figura 4.7 podemos observar uma comparação entre a função de distribuição empírica e a função de distribuição estimada, além do QQ-plot dos dados para a distribuição gama generalizada. Podemos observar, especialmente na Figura 4.7 (b), fortes indícios de que os dados não sejam oriundos da distribuição Gama Generalizada.

Em decorrência do que observamos nas Figuras 4.4, 4.5, 4.6 e 4.7, acreditamos que os dados sejam oriundos de uma distribuição com cauda de variação regular. Dessa forma, pelos resultados obtidos na Seção 3.2, faremos uso de $k = 1,5(\log n)^2$ para o limiar $X_{n-k,n}$. Como o conjunto de dados é composto de $n = 1758$ elementos, usaremos aqui $k = 83$.

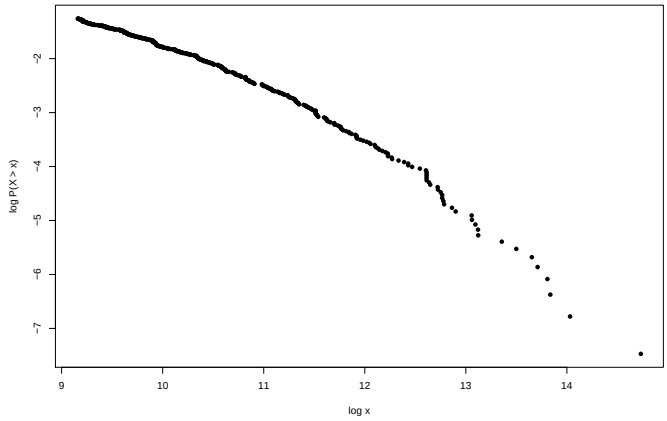

Figura 4.4 *Gráfico log-log aplicado aos dados referentes ao pagamentos de indenizações de seguros de incêndio no Brasil em 2003.*

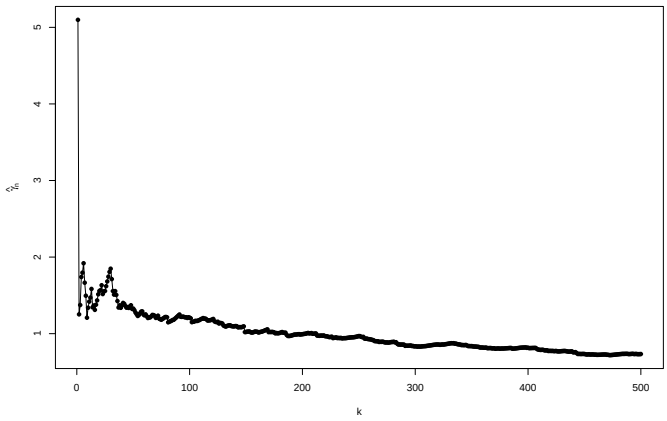

Figura 4.5 *Gráfico das estimativas de Hill aplicado aos dados referentes ao pagamentos de indenizações de seguros de incêndio no Brasil em 2003.*

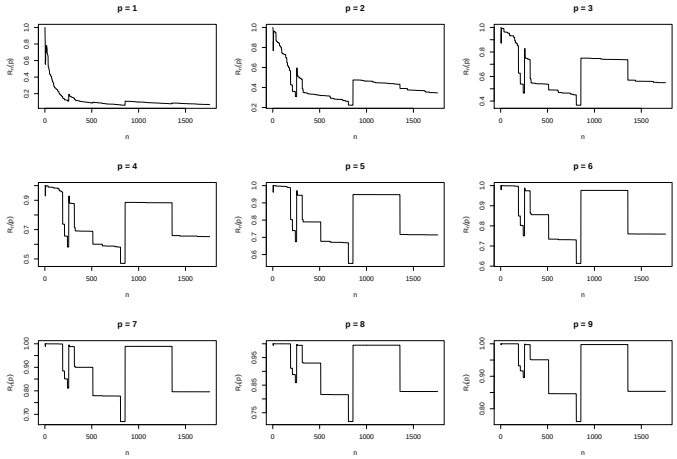

Figura 4.6 *Gráfico da razão entre máximos e somas aplicado aos dados referentes ao pagamentos de indenizações de seguros de incêndio no Brasil em 2003.*

Calculamos intervalos com 90% de confiança para os quantis de ordem $0,99$ e $0,999$, utilizando todos os métodos exibidos nesta dissertação, apesar de termos observado aqui fortes indícios de que tal conjunto de dados não possui distribuição Gama Generalizada. A finalidade de empregarmos o método da gama generalizada é apenas ilustrativa. Os intervalos de confiança encontram-se na Tabela 4.2, onde podemos observar que os intervalos de confiança obtidos sob os métodos da aproximação pela normal, razão de verossimilhanças e *data tilting* apresentam resultados parecidos entre si, e bastante diferentes do método da gama generalizada. Contudo, baseados nos mesmos resultados da Seção 3.3, acreditamos que o método da razão de verossimilhanças esteja fornecendo o melhor intervalo dentre todos os métodos utilizados aqui.

		$x_{0,01}$		$x_{0,001}$	
	k	Lim. Inf.	Lim. Sup.	Lim. Inf.	Lim. Sup.
Aprox. normal	83	303603,6	491874,3	1540976,3	5107601,2
Razão de Veross.	83	299038,3	532410,9	1617380,9	5649506,2
Data Tilting	83	301808,4	498909,3	1685386,8	4775026,0
Gama Generalizada		209973,8	271420,2	668531,2	904782,6

Tabela 4.2 *Intervalos de 90% de confiança para os quantis de ordem* $0,99$ *e* $0,999$ *das indenizações dos seguros de incêndios.*

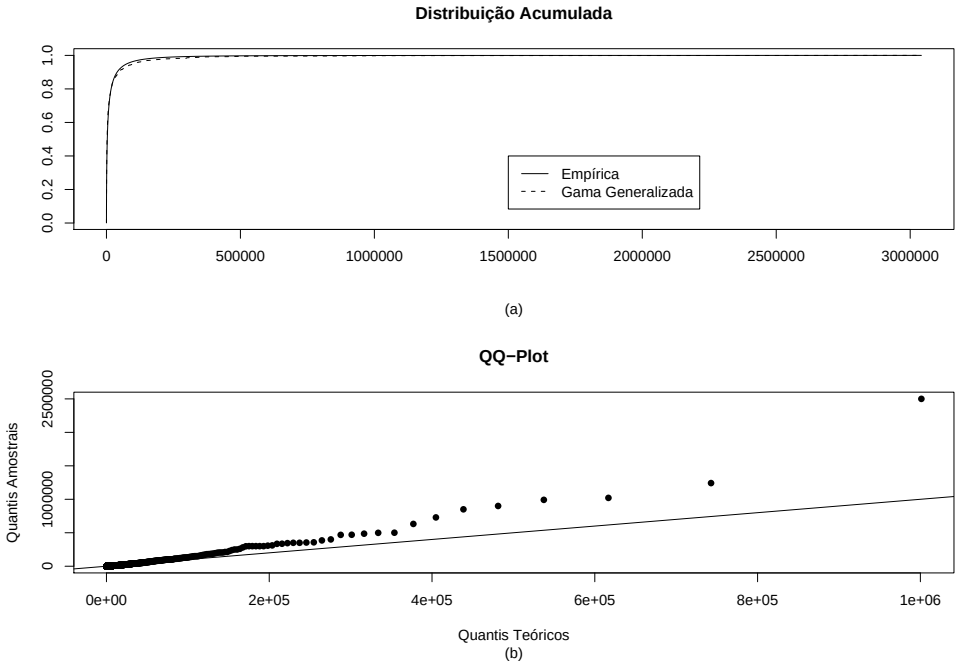

Figura 4.7 *(a) Gráfico da distribuição empírica com a função de distribuição ajustada pela gama generalizada e (b) QQ-plot dos dados comparados à gama generalizada*

Capítulo 5

Conclusões e perspectivas

Neste trabalho utilizamos os métodos da aproximação pela distribuição normal, razão de verossimilhanças, *data tilting* e gama generalizada a fim de obter intervalos de confiança para altos quantis oriundos de distribuições de caudas pesadas.

Observamos, via simulações de Monte Carlo, que o método da gama generalizada apresentou melhores intervalos de confiança para quantis de ordem $0,99$ e $0,999$, a um nível de 90%, no sentido de que os mesmos resultaram em maiores probabilidades de cobertura e menores amplitudes médias, para os dados gerados de uma distribuição Weibull de parâmetros $\beta = 0,3$ e $\alpha = 1$, sendo que esta é um caso particular da distribuição Gama Generalizada. Além disso, o método da razão de verossimilhanças apresentou melhores intervalos para dados gerados de uma distribuição Fréchet de parâmetro $\alpha = 1$.

Os limites de confiança para os métodos da razão de verossimilhanças e do *data tilting* foram obtidos, de forma aproximada, utilizando os resultados (2.20) e (2.26), respectivamente. Para o método da razão de verossimilhanças, escolhemos um valor inicial do quantil x_p, de tal forma que o valor de $l(x_p)$, em (2.18), fosse menor do que o α-ésimo quantil de uma χ^2 com 1 grau de liberdade, u_α. Daí, ficamos reduzindo (aumentando) o valor de x_p em passos pequenos, de tamanho 0,1, até que $l(x_p) > u_\alpha$. Assim, o último valor de x_p, onde $l(x_p) \leq u_\alpha$, era considerado o limite inferior (superior) do intervalo. Nas simulações e na aplicação, construímos intervalos com 90% de confiança para os quantis de ordem $0,99$ e $0,999$. Os intervalos de confiança para o método do *data tilting* foram obtidos de forma similar ao método da razão de verossimilhanças.

Os intervalos de confiança, em todos os métodos, foram calculados utilizando o programa estatístico R. Os códigos podem ser disponibilizados, a quem estiver interessado, via e-mail: `montoril@ime.usp.br`.

Este trabalho possui algumas possibilidades de aperfeiçoamentos que podem ser realizados em trabalhos futuros, dentre os quais podemos destacar:

- Escrever rotinas computacionais para estimar os parâmetros da distribuição Gama Gene-

ralizada pelo método dos momentos, a fim de utilizar como vetor inicial de parâmetros no processo iterativo da estimação por máxima verossimilhança.

- Realizar mais simulações, utilizando desta vez, outras distribuições, especialmente distribuições subexponenciais que não pertençam à classe de distribuições com cauda de variação regular, com o intuito de verificar, principalmente, se o método da gama generalizada é, de fato, mais adequado do que os outros métodos utilizados nesta dissertação, para a classe de distribuições subexponenciais.

- Comparar as formas utilizadas nesta dissertação para a escolha do limiar, nos métodos da aproximação pela normal, razão de verossimilhanças e *data tilting*, com outras propostas utilizadas na literatura.

- Ampliar as comparações entre as técnicas de estimação intervalar utilizando outros métodos, e de preferência, para outras classes de distribuições de caudas pesadas. Um dos métodos que podem ser trabalhados é conhecido pela sigla POT (*Peaks Over Threshold*). Mais detalhes com relação a este método podem ser obtidos em, por exemplo, Embrechts *et al.* (1997). Métodos de reamostragem, como Jackknife, também podem vir a ser uma alternativa interessante.

Apêndice A

Demonstrações

A.1 Cálculo de $\hat{\gamma}_n$ e $\hat{c}_n$

Utilizando a log-verossimilhança em (2.7), temos que o valor de c que maximiza $l(\gamma, c)$ será

$$\begin{aligned}
\left.\frac{\partial l(\gamma, c)}{\partial c}\right|_{c=\hat{c}_n} &= 0 \\
\frac{k}{\hat{c}_n} + (k-n)\frac{X_{n-k,n}^{-\gamma}}{1-\hat{c}_n X_{n-k,n}^{-\gamma}} &= 0 \\
\frac{k}{\hat{c}_n} &= (n-k)\frac{X_{n-k,n}^{-\gamma}}{1-\hat{c}_n X_{n-k,n}^{-\gamma}} \\
\hat{c}_n &= \frac{k}{n} X_{n-k,n}^{\gamma} \qquad \text{(A.1)}
\end{aligned}$$

Já o valor de γ que maximiza $l(\gamma, \hat{c}_n)$ será

$$\begin{aligned}
\left.\frac{\partial l(\gamma, \hat{c}_n)}{\partial \gamma}\right|_{\gamma=\hat{\gamma}_n} &= 0 \\
\frac{k}{\hat{\gamma}_n} - \sum_{i=1}^{k} \log X_{n-i+1,n} + (n-k)\hat{c}_n \frac{\log X_{n-k,n}}{X_{n-k,n}^{\gamma} - \hat{c}_n} &= 0 \\
\frac{k}{\hat{\gamma}_n} &= \sum_{i=1}^{k} \log X_{n-i+1,n} - k \log X_{n-k,n} \\
\hat{\gamma}_n &= \left\{ \frac{1}{k} \sum_{i=1}^{k} (\log X_{n-i+1,n} - \log X_{n-k,n}) \right\}^{-1}
\end{aligned}$$

A.2 Relação entre as funções de distribuição e quantil

Como $A(t) \to 0$ quando $t \to \infty$ e o limite em (2.11) é finito, então

$$\lim_{t\to\infty} \frac{U(tx)}{U(t)} = x^{1/\gamma}, \quad x > 0, \qquad \text{(A.2)}$$

o que nos garante que U é uma função de variação regular, podendo assim ser escrita na forma

$$U(x) = \ell_U(x)x^{1/\gamma}, \quad x > 0, \tag{A.3}$$

onde ℓ_U é uma função de variação lenta.

Sem perda de generalidade podemos reescrever (A.3) como

$$U(1/p) = \ell_U(1/p)p^{-1/\gamma}, \tag{A.4}$$

onde o interesse agora é observar o comportamento de U quando $p \to 0$.

Fazendo $p = \bar{F}(y)$, e lembrando que $U(x) = \left(\frac{1}{1-F}\right)^{-1}(x)$, teremos

$$\bar{F}(y) = \underbrace{\{\ell_U(1/\bar{F}(y))\}^{\gamma}}_{\ell_F(y)} y^{-\gamma}.$$

Podemos escrever $\dfrac{1}{\bar{F}(x)} = g(x)$, sendo g, pelas propriedades da função de distribuição, uma função monótona não-decrescente com $\lim\limits_{x\to\infty} g(x) = \infty$.

Como ℓ_U é uma função de variação lenta e g é monótona não-decrescente com $\lim\limits_{x\to\infty} g(x) = \infty$, temos que $\ell_U(g(x)) = \ell(x)$ também é uma função de variação lenta, o que garante que $\ell_F(x)$ também será, pois

$$\lim_{t\to\infty} \frac{\ell_F(tx)}{\ell_F(t)} = \lim_{t\to\infty} \left\{\frac{\ell(tx)}{\ell(t)}\right\}^{\gamma} = \left\{\lim_{t\to\infty} \frac{\ell(tx)}{\ell(t)}\right\}^{\gamma} = 1,$$

ficando assim verificado que (2.11) implica (2.1).

A.3 Multiplicadores de Lagrange para obter γ e c no método da razão de verossimilhanças

O objetivo aqui é de maximizar $l(\gamma, c)$ sujeito às seguintes restrições

$$\gamma > 0, \quad c > 0, \quad \gamma \log x_p + \log\left(\frac{p_n}{c}\right) = 0 \tag{A.5}$$

Pelo método dos multiplicadores de Lagrange temos

$$g(\gamma, c, \lambda) = l(\gamma, c) + \lambda\left(\gamma \log x_p + \log\left(\frac{p_n}{c}\right)\right) \tag{A.6}$$

Desta forma, o valor $\bar{c}(\lambda)$ que maximiza a função $g(\gamma, c, \lambda)$ será

$$\begin{aligned}
\left.\frac{\partial g(\gamma, c, \lambda)}{\partial c}\right|_{c=\bar{c}(\lambda)} &= 0 \\
\frac{k}{\bar{c}(\lambda)} - (n-k)\frac{T^{-\gamma}}{1-cT^{-\gamma}} - \frac{\lambda}{\bar{c}(\lambda)} &= 0 \\
\frac{k-\lambda}{\bar{c}(\lambda)} &= \frac{n-k}{T^{\gamma} - \bar{c}(\lambda)} \\
\bar{c}(\lambda) &= T^{\gamma}\frac{k-\lambda}{n-\lambda} \qquad \text{(A.7)}
\end{aligned}$$

O valor $\bar{\gamma}(\lambda)$ que maximiza $g(\gamma, \bar{c}(\lambda), \lambda)$ será

$$\begin{aligned}
\left.\frac{\partial g(\gamma, \bar{c}(\lambda), \lambda)}{\partial c}\right|_{\gamma=\bar{\gamma}(\lambda)} &= 0 \\
\frac{k}{\bar{\gamma}(\lambda)} - \sum_{i=1}^{k} \log X_{n-i+1,n} + (n-k)\frac{\bar{c}(\lambda)T^{-\bar{\gamma}(\lambda)}\log T}{1-cT^{-\bar{\gamma}(\lambda)}} + \lambda \log x_p &= 0 \\
\frac{k}{\bar{\gamma}(\lambda)} - \sum_{i=1}^{k} \log X_{n-i+1,n} + (n-k)\frac{\frac{k-\lambda}{n-\lambda}\log T}{1-\frac{k-\lambda}{n-\lambda}} + \lambda \log x_p &\overset{(A.7)}{=} 0 \\
\frac{k}{\bar{\gamma}(\lambda)} - \sum_{i=1}^{k} \log X_{n-i+1,n} + (k-\lambda)\log T + \lambda \log x_p &= 0 \\
\bar{\gamma}(\lambda) = \frac{k}{\sum_{i=1}^{k} \log \frac{X_{n-i+1,n}}{T} + \lambda(\log T + \log x_p)} & \qquad \text{(A.8)}
\end{aligned}$$

onde λ satisfaz

$$\bar{\gamma}(\lambda)\log x_p + \log\left(\frac{p_n}{\bar{c}(\lambda)}\right) = 0, \qquad \text{(A.9)}$$

$$\bar{\gamma}(\lambda) > 0 \quad \text{e} \quad \bar{c}(\lambda) > 0. \qquad \text{(A.10)}$$

A.4 Cálculo dos estimadores de γ e c no método *data tilting*.

Temos o interesse em obter

$$(\hat{\gamma}(\boldsymbol{q}), \hat{c}(\boldsymbol{q})) = \arg\max_{(\gamma,c)} \sum_{i=1}^{n} q_i \log \left\{ (c\gamma X_i^{-\gamma-1})^{\delta_i} (1 - cX_{n-k,n}^{-\gamma})^{1-\delta_i} \right\}$$

Inicialmente, denotemos

$$f(\gamma, c) = \sum_{i=1}^{n} q_i \log \left\{ (c\gamma X_i^{-\gamma-1})^{\delta_i} (1 - cX_{n-k,n}^{-\gamma})^{1-\delta_i} \right\}.$$

Não é difícil obter que

$$f(\gamma, c) = \sum_{i=1}^{n} q_i \delta_i \log c + \sum_{i=1}^{n} q_i \delta_i \log \gamma - (\gamma + 1) \sum_{i=1}^{n} q_i \delta_i \log X_i + \sum_{i=1}^{n} q_i (1 - \delta_i) \log (1 - cX_{n,n-k}).$$

Portanto, o valor de c que maximiza $f(\gamma, c)$ será

$$\begin{aligned}
\left. \frac{\partial f(\gamma, c)}{\partial c} \right|_{c=\hat{c}(\boldsymbol{q})} &= 0 \\
\frac{1}{\hat{c}(\boldsymbol{q})} \sum_{i=1}^{n} q_i \delta_i &= \frac{X_{n,n-k}^{-\gamma}}{1 - \hat{c}(\boldsymbol{q}) X_{n,n-k}^{-\gamma}} \sum_{i=1}^{n} q_i (1 - \delta_i) \\
\sum_{i=1}^{n} q_i \delta_i - \hat{c}(\boldsymbol{q}) X_{n,n-k}^{-\gamma} \sum_{i=1}^{n} q_i \delta_i &= \hat{c}(\boldsymbol{q}) X_{n,n-k}^{-\gamma} \sum_{i=1}^{n} q_i (1 - \delta_i) \\
\hat{c}(\boldsymbol{q}) &= X_{n,n-k}^{\gamma} \sum_{i=1}^{n} q_i \delta_i = X_{n,n-k}^{\hat{\gamma}(\boldsymbol{q})} \sum_{i=1}^{n} q_i \delta_i. \qquad \text{(A.11)}
\end{aligned}$$

Já o valor de γ que maximiza $f(\gamma, \hat{c}(\boldsymbol{q}))$ será

$$\begin{aligned}
\left. \frac{\partial f(\gamma, \hat{c}(\boldsymbol{q}))}{\partial \gamma} \right|_{\gamma=\hat{\gamma}(\boldsymbol{q})} &= 0 \\
\frac{1}{\hat{\gamma}(\boldsymbol{q})} \sum_{i=1}^{n} q_i \delta_i &= \sum_{i=1}^{n} q_i \log X_i - \sum_{i=1}^{n} q_i (1 - \delta_i) \frac{\hat{c}(\boldsymbol{q}) X_{n,n-k}^{-\hat{\gamma}(\boldsymbol{q})} \log X_{n,n-k}}{1 - \hat{c}(\boldsymbol{q}) X_{n,n-k}^{-\hat{\gamma}(\boldsymbol{q})}} \\
\frac{1}{\hat{\gamma}(\boldsymbol{q})} \sum_{i=1}^{n} q_i \delta_i &\overset{(A.11)}{=} \sum_{i=1}^{n} q_i \log X_i - \sum_{i=1}^{n} q_i \log X_{n,n-k} \\
\hat{\gamma}(\boldsymbol{q}) &= \frac{\sum_{i=1}^{n} q_i \delta_i}{\sum_{i=1}^{n} q_i \log (X_i - X_{n,n-k})}.
\end{aligned}$$

A.5 Multiplicadores de Lagrange para o vetor de pesos do método *data tilting*

$$D_\rho(\boldsymbol{q}) = \begin{cases} (\rho(1-\rho))^{-1}\left\{1 - n^{-1}\sum_{i=1}^n (nq_i)^\rho\right\}, & \text{se } \rho \neq 0, 1; \\ -n^{-1}\sum_{i=1}^n \log(nq_i), & \text{se } \rho = 0; \\ \sum_{i=1}^n q_i \log(nq_i), & \text{se } \rho = 1. \end{cases}$$

O objetivo aqui é de minimizar $D_\rho(\mathbf{q})$ sujeito às seguintes restrições

$$\boldsymbol{q} \geq \mathbf{0}; \quad \sum_{i=1}^n q_i = 1; \quad \hat{\gamma}(\boldsymbol{q}) \log \frac{x_p}{T} = \log \frac{\sum_{i=1}^n q_i \delta_i}{p_n} \tag{A.12}$$

onde $\mathbf{0} = (0, 0, \ldots, 0)^\top$, $\hat{\gamma}(\boldsymbol{q}) = \frac{\sum_{i=1}^n q_i \delta_i}{\sum_{i=1}^n q_i \delta_i \log \frac{X_i}{T}}$ e $T = X_{n,n-k}$.

Pelo método dos multiplicadores de Lagrange temos (no geral)

$$f(\boldsymbol{q}; \boldsymbol{\lambda}) = D_\rho(\boldsymbol{q}) + \lambda_1 (\sum_{i=1}^n q_i - 1) + \lambda_2 (\hat{\gamma}(\boldsymbol{q}) \log \frac{x_p}{T} - \log \frac{\sum_{i=1}^n q_i \delta_i}{p_n}) \tag{A.13}$$

Independente do valor de ρ, sempre teremos

$$\frac{\partial f(\boldsymbol{q}; \boldsymbol{\lambda})}{\partial \lambda_1} = \sum_{i=1}^n q_i - 1 = 0 \Leftrightarrow \sum_{i=1}^n q_i = 1 \tag{A.14}$$

$$\begin{aligned} \frac{\partial f(\boldsymbol{q}; \boldsymbol{\lambda})}{\partial \lambda_2} &= \hat{\gamma}(\boldsymbol{q}) \log \frac{x_p}{T} - \log \frac{\sum_{i=1}^n q_i \delta_i}{p_n} = 0 \\ &\Leftrightarrow \sum_{i=1}^n q_i \delta_i \log \frac{X_i}{T} = \frac{\sum_{i=1}^n q_i \delta_i \log \frac{x_p}{T}}{\log \frac{\sum_{i=1}^n q_i \delta_i}{p_n}} \end{aligned} \tag{A.15}$$

A partir de agora obteremos o valor de q_i $(j = 1, 2, \ldots, n)$ para cada um dos três casos de $D_\rho(\boldsymbol{q})$.

Quando $\rho \neq 0, 1$

$$\begin{aligned}
\frac{\partial f(\boldsymbol{q};\boldsymbol{\lambda})}{\partial q_j} &= -\frac{(nq_j)^{\rho-1}}{1-\rho} + \lambda_1 + \delta_j\lambda_2\left\{\left[\frac{\sum_{i=1}^n q_i\delta_i \log\frac{X_i}{T} - \sum_{i=1}^n q_i\delta_i \log\frac{X_j}{T}}{\left(\sum_{i=1}^n q_i\delta_i \log\frac{X_i}{T}\right)^2}\right]\log\frac{x_p}{T} - \frac{1}{\sum_{i=1}^n q_i\delta_i}\right\} \\
&= -\frac{(nq_j)^{\rho-1}}{1-\rho} + \lambda_1 + \delta_j\lambda_2\left\{\left[1 - \frac{\sum_{i=1}^n q_i\delta_i \log\frac{X_j}{T}}{\sum_{i=1}^n q_i\delta_i \log\frac{X_i}{T}}\right]\frac{\log\frac{x_p}{T}}{\sum_{i=1}^n q_i\delta_i \log\frac{X_i}{T}} - \frac{1}{\sum_{i=1}^n q_i\delta_i}\right\} \\
&\overset{(A.15)}{=} -\frac{(nq_j)^{\rho-1}}{1-\rho} + \lambda_1 + \delta_j\lambda_2\left\{\left[1 - \frac{\log\frac{\sum_{i=1}^n q_i\delta_i}{p_n}\log\frac{X_j}{T}}{\log\frac{x_p}{T}}\right]\frac{\log\frac{\sum_{i=1}^n q_i\delta_i}{p_n}}{\sum_{i=1}^n q_i\delta_i} - \frac{1}{\sum_{i=1}^n q_i\delta_i}\right\} \\
&= -\frac{(nq_j)^{\rho-1}}{1-\rho} + \lambda_1 + \delta_j\frac{\lambda_2}{\sum_{i=1}^n q_i\delta_i}\left\{\left[1 - \frac{\log\frac{\sum_{i=1}^n q_i\delta_i}{p_n}\log\frac{X_j}{T}}{\log\frac{x_p}{T}}\right]\log\frac{\sum_{i=1}^n q_i\delta_i}{p_n} - 1\right\} \quad \text{(A.16)}
\end{aligned}$$

Note que, de (A.14)

$$\sum_{i=1}^n q_i = \sum_{i=1}^n q_i(\delta_i + 1 - \delta_i) = \sum_{i=1}^n q_i\delta_i + \sum_{i=1}^n q_i(1-\delta_i) = 1 \Leftrightarrow \sum_{i=1}^n q_i\delta_i = 1 - \sum_{i=1}^n q_i(1-\delta_i)$$

Daqui por diante denotaremos $A_1(\lambda_1) = \sum_{i=1}^n q_i\delta_i$. Mostraremos que $A_1(\lambda_1)$ não depende de $\boldsymbol{q}$. Concluindo, como $\frac{\partial f(\boldsymbol{q};\boldsymbol{\lambda})}{\partial q_j} = 0$, de (A.16), temos que

$$q_j = \frac{1}{n}\left\{\lambda_1 + \delta_j\frac{\lambda_2}{A_1(\lambda_1)}\left\{\left[1 - \frac{A_1(\lambda_1)\log\frac{X_j}{T}}{\log\frac{x_p}{T}}\right]\log\frac{A_1(\lambda_1)}{p_n} - 1\right\}(1-\rho)\right\}^{\frac{1}{\rho-1}} \quad \text{(A.17)}$$

Assim, sendo $\delta_j = 0$, teremos

$$q_j = \frac{1}{n\lambda_1} \Rightarrow \sum_{i=1}^n q_i(1-\delta_i) = \frac{n-k}{n\lambda_1} \Rightarrow A_1(\lambda_1) = 1 - \frac{n-k}{n}\{(1-\rho)\lambda_1\}^{\frac{1}{\rho-1}}$$

Quando $\rho = 0$

Utilizando alguns resultados já obtidos anteriormente

$$\begin{aligned}
\frac{\partial f(\boldsymbol{q};\boldsymbol{\lambda})}{\partial q_j} &= -\frac{1}{nq_j} + \lambda_1 + \delta_j\frac{\lambda_2}{A_1(\lambda_1)}\left\{\left[1 - \frac{\log\frac{A_1(\lambda_1)}{p_n}\log\frac{X_j}{T}}{\log\frac{x_p}{T}}\right]\log\frac{A_1(\lambda_1)}{p_n} - 1\right\} \\
\Leftrightarrow q_j &= \frac{1}{n}\left\{\lambda_1 + \delta_j\frac{\lambda_2}{A_1(\lambda_1)}\left\{\left[1 - \frac{A_1(\lambda_1)\log\frac{X_j}{T}}{\log\frac{x_p}{T}}\right]\log\frac{A_1(\lambda_1)}{p_n} - 1\right\}\right\}^{-1} \quad \text{(A.18)}
\end{aligned}$$

Aqui teremos que $A_1(\lambda_1) = 1 - \frac{n-k}{n\lambda_1}$

Quando $\rho = 1$

$$\begin{aligned}\frac{\partial f(\boldsymbol{q};\boldsymbol{\lambda})}{\partial q_j} &= \log(nq_j) + 1 + \lambda_1 + \delta_j \frac{\lambda_2}{A_1(\lambda_1)} \left\{ \left[1 - \frac{\log \frac{A_1(\lambda_1)}{p_n} \log \frac{X_j}{T}}{\log \frac{x_p}{T}} \right] \log \frac{A_1(\lambda_1)}{p_n} - 1 \right\} = 0 \\ \Leftrightarrow q_j &= \frac{1}{n} \exp \left\{ -1 - \lambda_1 - \delta_j \frac{\lambda_2}{A_1(\lambda_1)} \left\{ \left[1 - \frac{A_1(\lambda_1) \log \frac{X_j}{T}}{\log \frac{x_p}{T}} \right] \log \frac{A_1(\lambda_1)}{p_n} - 1 \right\} \right\} \end{aligned} \quad (A.19)$$

Finalmente, aqui teremos que $A_1(\lambda_1) = 1 - \frac{n-k}{n} \exp\{-1 - \lambda_1\}$

A.6 Relação entre as distribuições Gama e Gama Generalizada

Inicialmente definamos X e Y, onde $X \sim \text{GG}(\beta, \zeta, \alpha)$ e $Y \sim \text{Gama}(\zeta, \alpha)$. Como já havíamos comentado, é possível escrever X como função de Y, e vice-versa, sob a seguinte relação

$$X = \alpha \left(\frac{Y}{\alpha} \right)^{\frac{1}{\beta}}.$$

Observe que

$$x = \alpha \left(\frac{y}{\alpha} \right)^{\frac{1}{\beta}} \Rightarrow y = \alpha \left(\frac{x}{\alpha} \right)^{\beta} \Rightarrow \frac{dy}{dx} = \beta \left(\frac{x}{\alpha} \right)^{\beta - 1}$$

Assim, pelo método do Jacobiano teremos

$$\begin{aligned} f_X(x) &= \left| \frac{dy}{dx} \right| f_Y(\alpha \left(\frac{x}{\alpha} \right)^{\beta}) && (A.20) \\ &= \beta \left(\frac{x}{\alpha} \right)^{\beta - 1} \frac{\beta \left[\alpha \left(\frac{x}{\alpha} \right)^{\beta} \right]^{\beta\zeta - 1}}{\Gamma(\zeta) \alpha^{\beta\zeta}} \exp \left\{ - \left(\frac{\alpha \left(\frac{x}{\alpha} \right)^{\beta}}{\alpha} \right)^{\beta} \right\} \mathbb{1}_{(0,\infty)}(x) && (A.21) \\ &= \frac{\beta x^{\beta\zeta - 1}}{\Gamma(\zeta) \alpha^{\beta\zeta}} \exp \left\{ - \left(\frac{x}{\alpha} \right)^{\beta} \right\} \mathbb{1}_{(0,\infty)}(x) && (A.22) \end{aligned}$$

Dessa forma temos que há relação entre o $(1-p)$-ésimo quantil da Gama Generalizada, x_p, e o $(1-p)$-ésimo quantil da Gama, y_p. Sejam F_X e F_y as funções de distribuição acumulada de X e Y, respectivamente (com X e Y definidos anteriormente). Assim, temos que o quantil de ordem

$(1-p)$ da Gama Generalizada satisfaz $F_X(x_p) = 1-p$ e, analogamente, o quantil de ordem $(1-p)$ da Gama satisfaz $F_Y(y_p) = 1-p$. Então

$$F_X(x_p) = \mathbb{P}(X \leq x_p) = \mathbb{P}\left(\alpha\left(\frac{Y}{\alpha}\right)^{\frac{1}{\beta}} \leq x_p\right) = \mathbb{P}\left(Y \leq \alpha\left(\frac{x_p}{\alpha}\right)^{\beta}\right) = F_Y\left(\alpha\left(\frac{x_p}{\alpha}\right)^{\beta}\right) = 1-p$$

Logo temos que

$$y_p = \alpha\left(\frac{x_p}{\alpha}\right)^{\beta} \Rightarrow x_p = \alpha\left(\frac{y_p}{\alpha}\right)^{\frac{1}{\beta}}.$$

Apêndice B

Método para a seleção do valor de k em $X_{n-k,n}$

B.1 Pelo erro quadrático médio assintótico do estimador de Hill

Falaremos aqui, sobre o método proposto por Beirlant *et al.* (2002), para obter o valor de k que minimiza o erro quadrático médio assintótico do estimador de Hill, o qual será definido aqui por $H_{k,n}$. Existe uma variedade de métodos que fazem tal tarefa, mas optamos por este devido a sua popularidade entre os estatísticos, segundo Beirlant *et al.* (2004), que também comenta alguns outros métodos.

Sejam $X_{1,n}, X_{2,n}, \ldots, X_{n,n}$ as estatísticas de ordem de uma amostra aleatória $X_1, X_2, \ldots, X_n$, independentes e com distribuição com cauda de varição regular do tipo

$$1 - F(x) = x^{-1/\gamma}\ell_F(x),$$

onde ℓ_F é uma função de variação lenta, como em (1.5), ou equivalentemente

$$\lim_{x\to\infty} \log \frac{\ell_F(\lambda x)}{\ell_F(x)} = 0, \quad \text{para todo} \quad \lambda > 0.$$

Para especificar a taxa em que tal limite é atendido, é necessária a seguinte suposição.

Suposição (R_ℓ): Existe uma constante real negativa, $\rho < 0$, e uma função de taxa, b, satisfazendo $b(x) \to \infty$ quando $x \to \infty$, tal que para todo $\lambda \geq 1$, quando $x \to \infty$

$$\log \frac{\ell(\lambda x)}{\ell(x)} \sim b(x)k_\rho(\lambda),$$

com $k_\rho(\lambda) = \dfrac{\lambda^\rho - 1}{\rho}$.

Tal suposição, segundo Matthys e Beirlant (2001), é válida para a maioria das distribuições (sob a definição utilizada nesta dissertação) com cauda de varição regular e, a partir de tal suposição, pode-se verificar que o estimador de Hill, $H_{k,n}$, é assintoticamente normal (ver Haeusler e Teugels

(1985)), para seqüências intermediárias de valores de k (i.e. $n \to \infty$, $k \to \infty$ e $\frac{k}{n} \to 0$) que converge para infinito em uma taxa apropriada:

$$\sqrt{k}\left(H_{k,n} - \gamma - \frac{b_{n,k}}{1-\rho}\right) \xrightarrow{D} N(0, \gamma^2),$$

onde $b_{n,k} = b\left(\frac{n+1}{k+1}\right)$ e $H_{k,n} = \frac{1}{k}\sum_{i=1}^{k}(\log X_{n-i+1,n} - \log X_{n-k,n})$.

Baseados nisso, temos que o viés assintótico (VA) e a variância assintótica (VarA) do estimador de Hill são, respectivamente

$$\mathrm{VA}(H_{k,n}) = \frac{b_{n,k}}{1-\rho}$$

e

$$\mathrm{VarA}(H_{k,n}) = \frac{\gamma^2}{k}.$$

Portanto, o erro quadrático médio assintótico (EQMA) do estimador de Hill será

$$\mathrm{EQMA}(H_{k,n}) = \mathrm{VA}^2(H_{k,n}) + \mathrm{VarA}(H_{k,n}) = \left(\frac{b_{n,k}}{1-\rho}\right)^2 + \frac{\gamma^2}{k}.$$

O problema que surge é que os valores γ, $b_{n,k}$ e ρ são desconhecidos na prática, sendo portanto, necessário estimá-los.

A estimação conjunta dos três parâmetros pode ser feita via máxima verossimilhança, utilizando o resultado que será comentado a seguir.

Seja $Z_j = j(\log X_{n-j+1,n} - \log X_{n-j,n}),\ j = 1, 2, \ldots, k$. É possível verificar a seguinte aproximação (ver Beirlant *et al.* (1999))

$$Z_j \sim \left(\gamma + b_{n,k}\left(\frac{j}{k+1}\right)^{-\rho}\right) E_j, \quad 1 \le j \le k,$$

onde $(E_1, E_2, \ldots, E_k)^\top$ é um vetor de variáveis aleatórias independentes, todas com distribuição exponencial de média 1. Dessa forma, não é difícil checar que, por aproximação, Z_j também possui distribuição exponencial, mas com média $\left(\gamma + b_{n,k}\left(\frac{j}{k+1}\right)^{-\rho}\right)$.

Com base nisso, é possível montar a função de verossimilhança tratando Z_j $(j = 1, 2, \ldots, k)$ como sendo uma exponencial e, assim, obter a seguinte log-verossimilhança para $Z_1, Z_2, \ldots, Z_k$,

$$l(\gamma, b_{n,k}, \rho) = -\sum_{j=1}^{k} \log\left\{\left(\gamma + b_{n,k}\left(\frac{j}{k+1}\right)^{-\rho}\right)\right\} - \sum_{j=1}^{k} \frac{Z_j}{\left(\gamma + b_{n,k}\left(\frac{j}{k+1}\right)^{-\rho}\right)}$$

da qual é possível estimar, iterativamente, os parâmetros γ, $b_{n,k}$ e ρ.

No sitio `http://lstat.kuleuven.be/Wiley` existe a rotina pronta, em código S-Plus, a qual pode ser utilizada também no software R. Tal endereço é referente a Beirlant *et al.* (2004).

Referências Bibliográficas

Adlouni, S. E., Bobée, B. e Ouarda, T. B. M. J. (2008). On the tails of extreme event distributions in hydrology. *Journal of Hydrology 355*(1-4), 16–33.

Ashkar, F. e Ouarda, T. B. M. J. (1998). Approximate confidence intervals for quantiles of gamma and generalized gamma distributions. *Journal of Hydrologic Engineering 3*(1), 43–51.

Beirlant, J., Dierckx, G., Goegebeur, Y. e Matthys, G. (1999). Tail index estimation and an exponential regression model. *Extremes 2*(2), 177–200.

Beirlant, J., Dierckx, G., Guillou, A. e Staăricaă, C. (2002). On exponential representations of log-spacings of extreme order statistics. *Extremes 5*(2), 157–180.

Beirlant, J., Goegebeur, Y., Segers, J. e Teugels, J. (2004). *Statistics of Extremes: Theory and Applications*. Wiley.

Casella, G. e Berger, R. (2001). *Statistical Inference* (2 ed.). Duxbury Resource Center.

Chan, N. H., Deng, S., Peng, L. e Xia, Z. (2007). Interval estimation of value-at-risk based on garch models with heavy-tailed innovations. *Journal of Econometrics 137*(2), 556–576.

Coles, S. (2001). *An Introduction to Statistical Modeling of Extreme Values*. Springer.

De Haan, L. e Stadtmüller, U. (1996). Generalized regular variation of second order. *J. Austral. Math. Soc. Ser. A 61*(3), 381–395.

Dudewicz, E. J. e Mishra, S. N. (1988). *Modern mathematical statistics* (2 ed.). Wiley.

Embrechts, P., Klüppelberg, C. e Mikosch, T. (1997). *Modelling extremal events*, Volume 33 of *Applications of Mathematics (New York)*. Berlin: Springer-Verlag.

Ferreira, A., De Haan, L. e Peng, L. (2003). On optimising the estimation of high quantiles of a probability distribution. *Statistics 37*(5), 401–434.

Goldie, C. e Kluppelberg, C. (1997). Subexponential distributions.

Haeusler, E. e Teugels, J. L. (1985). On asymptotic normality of hill's estimator for the exponent of regular variation. *Ann. Statist. 13*(2), 743–756.

Hall, P. e Yao, Q. (2003). Data tilting for time series. *J. R. Stat. Soc. Ser. B Stat. Methodol. 65*(2), 425–442.

Hill, B. M. (1975). A simple general approach to inference about the tail of a distribution. *Ann. Statist. 3*(5), 1163–1174.

Kitamura, Y. (2006). Empirical likelihood methods in econometrics: Theory and practice. Technical Report 1569.

Lehmann, E. L. e Casella, G. (2003). *Theory of Point Estimation*. Springer.

Matthys, G. e Beirlant, J. (2001). Extreme quantile estimation for heavy-tailed distributions.

Matthys, G., Delafosse, E., Guillou, A. e Beirlant, J. (2004). Estimating catastrophic quantile levels for heavy-tailed distributions. *Insurance: Mathematics and Economics 34*(3), 517–537.

Melo, E. F. L. (2006). Uma aplicacao da teoria de valores extremos para avaliacao do risco de contratos de resseguro. *Revista Brasileira de Risco e Seguro 2*(3), 1–22.

Mood, A. M., Graybill, F. A. e Boes, D. C. (1974). *Introduction to the Theory of Statistics* (3 ed.). McGraw-Hill Companies.

Nelder, J. A. e Mead, R. (1965). A simplex method for function minimization. *The Computer Journal 7*(4), 308–313.

Owen, A. B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika 75*(2), 237–249.

Owen, A. B. (2001). *Empirical Likelihood*. New York: Chapman & Hall/CRC.

Peng, L. e Qi, Y. (2006). Confidence regions for high quantiles of a heavy tailed distribution. *Ann. Statist. 34*(4), 1964–1986.

R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.

Rust, H. W., Kallache, M., j. Schellnhuber, H. e Kropp, J. P. (2006). Confidence intervals for flood return level estimates using a bootstrap approach.

Schervish, M. J. (1996). *Theory of Statistics*. Springer.

Sigman, K. (1999). A primer on heavy-tailed distributions. *Queueing Syst. Theory Appl. 33*(1-3), 261–275.

Stacy, E. W. (1962). A generalization of the gamma distribution. *Ann. Math. Statist. 33*(3), 1187–1192.

Stacy, E. W. e Mihram, G. A. (1965). Parameter estimation for a generalized gamma distribution. *Technometrics 7*(3), 349–358.

Toomet, O. e Henningsen, A. (2008). *maxLik: Maximum Likelihood Estimation*. R package version 0.5-2.

Tsay, R. S. (2002). *Analysis of Financial Time Series* (2nd ed.). Wiley.

Van Noortwijk, J. M. (1999). Quantiles of generalised gamma distributions from a bayesian point of view.

Yee, T. W. (2008). *VGAM: Vector Generalized Linear and Additive Models*. R package version 0.7-7.