REJANE DOS SANTOS BRITO

ESTUDO DE EXPANSÕES ASSINTÓTICAS, AVALIAÇÃO NUMÉRICA DE MOMENTOS DAS DISTRIBUIÇÕES BETA GENERALIZADAS, APLICAÇÕES EM MODELOS DE REGRESSÃO E ANÁLISE DISCRIMINANTE

RECIFE-PE - MAR/2009

UNIVERSIDADE FEDERAL RURAL DE PERNAMBUCO

PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO

PROGRAMA DE PÓS-GRADUAÇÃO EM BIOMETRIA E ESTATÍSTICA APLICADA

ESTUDO DE EXPANSÕES ASSINTÓTICAS, AVALIAÇÃO NUMÉRICA DE MOMENTOS DAS DISTRIBUIÇÕES BETA GENERALIZADAS, APLICAÇÕES EM MODELOS DE REGRESSÃO E ANÁLISE DISCRIMINANTE

Dissertação apresentada ao Programa de Pós-Graduação em Biometria e Estatística Aplicada como exigência parcial à obtenção do título de Mestre.

Área de Concentração: Modelagem Estatística e Computacional

Orientadora: Profa. Dra. Laélia Pumilla Botêlho Campos dos Santos

Co-orientador: Prof. Dr. Gauss Moutinho Cordeiro

RECIFE-PE - MAR/2009.

FICHA CATALOGRÁFICA

B862e Brito, Rejane dos Santos
Estudo de expansões assintóticas, avaliação numérica de momentos das distribuições beta generalizadas, aplicações em modelos de regressão e análise discriminante / Rejane dos Santos Brito. -- 2009.
103 f. : il.

Orientadora : Laélia Pumilla Botêlho C. dos Santos
Dissertação (Mestrado em Biometria e Estatística Aplica – da) -- Universidade Federal Rural de Pernambuco. Departa – mento de Estatística e Informática.
Inclui apêndice e bibliografia.

CDD 519. 54

1. Aproximação ponto de sela
2. Distribuições beta generalizadas
3. Distribuição beta power
4. Regressão logística
5. Análise discriminante
I. Santos, Laélia Pumilla Botelho Campos dos
II. Título

UNIVERSIDADE FEDERAL RURAL DE PERNAMBUCO

PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO

PROGRAMA DE PÓS-GRADUAÇÃO EM BIOMETRIA E ESTATÍSTICA APLICADA

ESTUDO DE EXPANSÕES ASSINTÓTICAS, AVALIAÇÃO NUMÉRICA DE MOMENTOS DAS DISTRIBUIÇÕES BETA GENERALIZADAS, APLICAÇÕES EM MODELOS DE REGRESSÃO E ANÁLISE DISCRIMINANTE

Rejane dos Santos Brito

Dissertação julgada adequada para obtenção do título de mestre em Biometria e Estatística Aplicada, defendida e aprovada por unanimidade em 20/03/2009 pela Comissão Examinadora.

Orientadora:

Laélia P. B. Campos dos Santos

Profa. Dra. Laélia P. B. Campos dos Santos
Universidade Federal Rural de Pernambuco

Banca Examinadora:

Cláudio Tadeu Cristino

Prof. Dr. Cláudio Tadeu Cristino
Universidade Federal Rural de Pernambuco

Audrey Helen M. de A. Cysneiros

Profa. Dra. Audrey Helen M. de Aquino Cysneiros
Universidade Federal de Pernambuco

Francisco José de Azevêdo Cysneiros

Prof. Dr. Francisco José de A. Cysneiros
Universidade Federal de Pernambuco

Dedico este trabalho

aos meus amados pais, Cicero e Creuza.

Agradecimentos

A realização deste trabalho só foi possível por graça de Deus,

por ter concedido a mim o dom de viver na determinação

da busca da realização de mais um sonho.

Meus sinceros agradecimentos...

...aos meus pais, Cicero e Maria Creuza, que com amor, compreensão e cumplicidade sempre me apoiaram e incentivaram em todos os acontecimentos de minha vida. À vocês todo meu respeito, admiração e amor;

...aos meus irmãos, Rauflan e Ronisson, que tanta força, incentivo, amor e carinho proporcionaram de forma a auxiliar a realização de mais um sonho;

...ao meu amado, Patrick Bager, por existir, pela alegria, força e amparo nos momentos delicados da minha vida. À você todo meu amor;

...aos demais familiares, que sempre me deram carinho, apoio e compreenderam minha ausência em determinados momentos;

...à minha orientadora, professora Dra. Laélia Pumilla Botêlho Campos dos Santos, pela orientação, apoio e amizade no desenvolvimento deste trabalho;

...ao meu co-orientador, professor Dr. Gauss Moutinho Cordeiro pela orientação, apoio, amizade, confiança, paciência e comprometimento no desenvolvimento desta dissertação;

...à empresa PETROBRAS pelo suporte sempre ofertado desde a minha jornada universitária cujo auxílio proporcionou o desenvolvimento de parte deste trabalho. A todos os que compõem a ST/CER. Um agradecimento especial a Marcelo Hardman e Vitor Hugo Simon pela amizade, incentivo e dedicação;

...aos colegas do mestrado pela amizade compartilhada;

...aos professores e funcionários do Departamento de Estatística e Informática pela dedicação e apoio, em especial a Marco Santos e Zuleide França;

...às amizades conquistadas durante minha estada em Recife. Em especial a Alessandro Santos, Amanda Lira, Artur Lemonte, Daniela Nava, Getúlio Amaral, Gleifer Vaz, Hemílio Fernandes, Jones Albuquerque, Leila Rameh, Líliam Medeiros, Luz Marina, Marcelo Rodrigo, Mariana Dantas, Munindra Mohan, Murilo Medeiros, Tadeu Rodrigues, Themis Abensur e Vanessa Kelly pela amizade, companheirismo e incentivo sem os quais não teria enfrentado os momentos difíceis dessa minha trajetória;

...aos participantes da banca examinadora pelas valiosas contribuições;

...à CAPES pelo suporte financeiro.

Enfim, a todos que de forma direta ou indireta, contribuíram para realização deste trabalho.

“All Knowledge is, in final analysis, History.
All sciences are, in the abstract, Mathematics.
All judgements are, in their rationale, Statistics.”

Radhakrishna Rao.

Resumo

Inicialmente, realiza-se uma revisão literária sobre as expansões assintóticas de Daniels, Edgeworth, Lugannani-Rice e Cordeiro-Ferrari. Mediante uso da expansão de Cordeiro-Ferrari, torna-se possível realizar um estudo correspondente a aproximação da distribuição gama $G(\mu, \phi)$ em função da distribuição exponencial com média α. E, ainda, numa outra aplicação, faz-se a aproximação da distribuição t-Student com ν graus de liberdade em função da distribuição normal padrão. Além disso, apresenta-se um estudo correspondente às funcionalidades das distribuições beta generalizadas e, ainda, a obtenção dos momentos das distribuições beta generalizadas mediante as funções de Lauricella e generalizada de Kampé de Fériet. Propõe-se, ainda, a generalização da distribuição power como sendo uma nova distribuição beta generalizada. Por fim, realizam-se algumas aplicações em modelos de regressão, mediante regressão logística, bem como em modelos de análise discriminante.

Palavras-chave: Aproximação Ponto de Sela, Distribuições Beta Generalizadas, Distribuição Beta Power, Regressão Logística, Análise Discriminante.

Abstract

We make a review about Edgeworth, Lugannani-Rice, Daniels and Cordeiro-Ferrari asymptotic approximations. We use the Cordeiro-Ferrari asymptotic approximation to approximate the gamma distribution $G(\mu, \phi)$ by the exponential distribution with mean α. In a further application, based on the statistical proposed by them, we approximate the t-Student distribution with ν degrees of freedom using the normal standard distribution. Moreover, we realize a study about the functionalities of the beta generalized distributions. We obtain moments of the generalized beta distributions using the Lauricella and Kampé de Fériet generalized functions. Beyond this, we propose a new generalized beta distribution called beta power. Finally, we realize some applications in regression models by logistic regression and further more using discriminant analysis.

Key words: Saddle Point Approximation, Generalized Beta Distribution, Beta Power Distribution, Logistic Regression, Discriminant Analysis.

Lista de Figuras

Lista de Tabelas

Sumário

1 Introdução

Inicialmente, propõe-se uma revisão literária referente a obtenção das expansões ponto de sela mediante as expansões assintóticas de Edgeworth, Lugannani-Rice, Daniels e Cordeiro-Ferrari. Em aplicações na Estatística, Daniels (1954) foi o pioneiro no estudo das expansões ponto de sela. Em muitos casos, a metodologia ponto de sela é utilizada para determinar limites uniformes com erros relativos sobre o intervalo da distribuição (KOLASSA, 1997). Segundo Goutis e Casella (1999), as expansões ponto de sela são muito importantes na teoria assintótica, pois aproximam com grande precisão as funções densidade e de distribuição da soma e da média de variáveis aleatórias. Reid (1988) fez uma derivação da expansão ponto de sela denominada de "versão mais estatística" baseada na expansão de Edgeworth.

Cordeiro e Ferrari (1998) propuseram uma estatística que aproxima a soma padronizada de variáveis aleatórias independentes e identicamente distribuídas até ordem $O(n^{-1})$, em que n é o tamanho da amostra. Diante disto, utilizando o conhecimento referente à correção de Bartlett generalizada definida por Cordeiro e Ferrari (1998), utiliza-se a expansão proposta por eles para aproximar a distribuição gama $G(\mu, \phi)$ em função da distribuição exponencial com média α. E, ainda, numa outra aplicação, faz-se a aproximação da distribuição t-Student com ν graus de liberdade em função da distribuição normal padrão (BRITO; CORDEIRO, 2009).

Realiza-se, ainda, um resumo literário das distribuições beta generalizadas com o intuito de mostrar as funcionalidades dessas distribuições, o cálculo dos seus momentos ordinários, bem como os gráficos correspondentes de assimetria e curtose (ver Eugene et al. (2002), Nadarajah e Kotz (2005), Barreto-Souza et al. (2009)).

Sabe-se que a distribuição beta é uma das mais usadas para modelar experimentos aleatórios que produzem resultados no intervalo $(0, 1)$, dada a grande flexibilidade de ajuste de seus parâmetros. Isto a torna a mais flexível da família de distribuições. Com o objetivo de ilustrar a versatilidade desta distribuição, faz-se uma aplicação desta em estudos agrários.

Supondo que as propriedades da $F(x)$ poderia, em princípio, seguir das propriedades da função hipergeométrica e sabendo, ainda, que as funções de Lauricella e de Kampé de Fériet são generalizações das funções hipergeométricas de Gauss, propõe-se o uso destas no cálculo dos momentos das distribuições beta generalizadas. Por conseguinte, faz-se o uso das funções de Lauricella mediante cálculo dos momentos de distribuições empíricas. E, no caso da função generalizada de Kampé de Fériet, obtém-se uma parametrização que proporciona o cálculo dos momentos da distribuição beta beta e uma segunda parametrização para o cálculo dos momentos da distribuição beta Student t.

Com o intuito de contribuir com mais uma nova distribuição é proposta uma generalização para a distribuição power, sendo esta nova distribuição denominada de Beta Power (BP). Para isso, encontra-se a forma analítica da sua função densidade de probabilidade e a função da razão de risco, calcula-se o n-ésimo momento e analisa-se a variação das medidas de assimetria e curtose; como também, faz-se a estimação dos parâmetros através do método de máxima verossimilhança. Para esta nova distribuição, realizam-se aplicações a dados reais e simulados.

Por fim, propõe-se um modelo que solucione o problema abordado em estudos geológicos referente a poços de petróleo. A justificativa para realização desse estudo surge da extrema necessidade das empresas petrolíferas obterem informações de tipos de rochas mediante a perfuração de poços, sendo as mesmas capazes de acumular petróleo ou não. Nesse contexto, para o modelo proposto, é feita uma comparação entre a metodologia usando regressão logística e análise discriminante

As análises desenvolvidas neste trabalho foram conduzidas usando as rotinas computacionais do software R, versão 2.8.0 para o sistema operacional Microsoft Windows. Detalhes sobre este ambiente de programação podem ser encontrados no endereço eletrônico www.r-project.org. Utilizaram-se, também, os softwares MAPLE® na versão 11.0, MATLAB® na versão 7.3(R2006b) e MATHEMATICA® na versão 5.0 para desenvolvimento das análises.

Essa dissertação está dividida em 5 capítulos, sendo que cada capítulo segue estudos estatísticos independentes. No primeiro capítulo é apresentada uma síntese desta dissertação e as justificativas para o desenvolvimento da mesma. O Capítulo 2 é baseado em teoria assintótica, onde é abordada algumas expansões ponto de sela. Os Capítulos 3 e 4 correspondem ao estudo das distribuições beta generalizadas, sendo que o Capítulo 4 propõe uma generalização da distribuição power. O Capítulo 5 trata de uma proposta de modelos para análise de dados de reservatório de petróleo.

2 Comparação de Expansões Assintóticas de Edgeworth, Lugannani e Rice, Daniels e Cordeiro-Ferrari

O estudo das expansões ponto de sela teve início com Esscher (1932), mas Daniels (1954) foi o pioneiro no assunto, pois aplicou a expansão ponto de sela na área da estatística. Lugannani e Rice (1980) apresentaram a expansão ponto de sela e sua aplicabilidade referente à soma estocástica de variáveis aleatórias i.i.d. (independentes e identicamente distribuídas).

Em muitas aplicações estatísticas, as expansões têm sua importância no que se refere, por exemplo, à construção de testes e intervalos de confiança, além, também, do cálculo dos p-valores. Em muitos casos, a metodologia ponto de sela é utilizada para determinar limites uniformes com erros relativos sobre o intervalo da distribuição (KOLASSA, 1997). Sabe-se ainda que as expansões ponto de sela são muito importantes na teoria assintótica, pois aproximam com grande precisão as funções densidade e de distribuição da soma e da média de variáveis aleatórias i.i.d. (GOUTIS; CASELLA, 1999). Uma escolha da aproximação de forma ampla é baseada na expansão de Edgeworth e na expansão inversa de Cornish-Fisher, mas, apesar da importância no estudo teórico das funções densidade e distribuição e aproximação dos quantis, estas expansões podem dar estimativas negativas da função densidade e, também, aproximações não-monótonas da função distribuição e das funções dos quantis quando aplicadas na prática. Um alternativa natural é o uso das ferramentas de expansão assintótica para integrais, tais como, aproximações de Laplace e aproximações ponto de sela (DAVISON, 2001).

A obtenção das expansões ponto de sela é feita através do uso de ferramentas como exponencial “inclinada”, expansões de Edgeworth, polinômios de Hermite, integrações complexas e outras noções avançadas. As expansões ponto de sela são facilmente deduzidas da função geratriz de cumulantes da variável aleatória de interesse.

2.1 Aproximação de Laplace

Goutis e Casella (1999) explicitam a obtenção da expansão de Laplace mediante o uso dos primeiros termos da expansão em série de Taylor. Utiliza-se a função $h(y) = \log f(y)$ tal que $f(y) = \exp\{h(y)\}$. Escolhendo $\hat{y}$ como um ponto a ser expandido em torno de y, tem-se

$$f(y) \approx \exp\left\{h(\hat{y}) + \frac{(y-\hat{y})^2}{2}h''(\hat{y})\right\}. \tag{2.1}$$

A aproximação (2.1) pode ser usada para calcular integrais de funções positivas, tais como $\int f(y)dy$ dentro do intervalo (a,b). Ao expandir o integrando como em (2.1), obtém-se

$$\int_a^b f(y)dy \approx \int_a^b \exp\left\{h(\hat{y}) + \frac{(y-\hat{y})^2}{2}h''(\hat{y})\right\}dy. \tag{2.2}$$

Se $\hat{y}$ é um ponto de máximo da função $f(y)$, $h''(\hat{y})$ é negativo e o lado direito de (2.2) pode ser calculado, pois a parte principal da equação representa o integrando da distribuição normal com média $\hat{y}$ e variância $-1/h''(\hat{y})$. Dessa forma, sendo agora o intervalo (a,b) equivalente ao intervalo $(-\infty,+\infty)$, obtém-se

$$\int_{-\infty}^{+\infty} f(y)dy \approx \exp\{h(\hat{y})\}\left\{-\frac{2\pi}{h''(\hat{y})}\right\}^{1/2}. \tag{2.3}$$

A equação (2.3) é conhecida como aproximação de Laplace para integrais.

2.1.1 Fórmula da Inversão e função geradora de cumulantes

Considere que Y é uma variável aleatória tendo função densidade ou função de probabilidade $f(y)$. As diferentes maneiras de aproximar estas funções são obtidas calculando a função geradora de momentos

$$\phi(\lambda) = \int_{-\infty}^{+\infty} \exp(\lambda y) f(y)dy. \tag{2.4}$$

A fórmula da transformação inversa implica

$$\begin{aligned} f(y) &= \frac{1}{2\pi}\int_{-\infty}^{+\infty} \phi(i\lambda)\exp(-i\lambda y)d\lambda \\ &= \frac{1}{2\pi}\int_{-\infty}^{+\infty} \exp\{K(i\lambda) - i\lambda y\}d\lambda, \end{aligned} \tag{2.5}$$

em que $i = \sqrt{-1}$, $\phi(i\lambda)$ é a função característica de Y e $K(\lambda) = \log\phi(\lambda)$. A função $K(\lambda)$ é conhecida como função geradora de cumulantes (f.g.c.) de Y e desempenha um papel importante na teoria assintótica.

2.2 Expansão ponto de sela

A primeira aplicação estatística da expansão ponto de sela foi feita por Daniels (1954) através da determinação de uma aproximação da função densidade utilizando a transformada de Fourier. A aproximação pode ser obtida através da família exponencial conjugada e, também, através de uma relação entre o ponto de sela e o estimador de máxima verossimilhança (EMV). Uma dificuldade em se usar esta expansão vem do fato de ser necessário o uso de integrações complexas.

Segundo Daniels (1954), existem formas distintas de se obter (analiticamente) a aproximação ponto de sela e suas expansões associadas em potências de ordem $O(n^{-1})$ para a função densidade da média amostral $\bar{Y}$ de uma amostra aleatória i.i.d.

2.2.1 Expansão ponto de sela através do método de Laplace

Considera-se a técnica da expansão de Laplace apresentada na Seção 2.1. Realizando uma mudança de variável na fórmula da transformação inversa (2.5), $\lambda' = i\lambda$, obtém-se uma função densidade em termos da integral complexa

$$f(y) = \frac{1}{2\pi i}\int_{t-i\infty}^{t+i\infty} \exp\{K(\lambda) - \lambda y\} d\lambda \tag{2.6}$$

para t em torno de zero. Segundo o Teorema da Curva Fechada[1], pode-se aproximar um valor de t de modo a obter a integração.

Aproximando a integral de $\exp\{K(\lambda)\}$ com relação à variável λ através da expansão de Laplace, obtém-se a equação

$$f(y) \approx \int_{-\infty}^{+\infty} \exp\left\{ K(\hat{\lambda}) + \frac{(\lambda-\hat{\lambda})^2}{2} \frac{d^2K(\lambda)}{d\lambda^2}\bigg|_{\hat{\lambda}} \right\} d\lambda$$

ou

$$f(y) \approx \exp\{K(\hat{\lambda})\} \left\{ -\frac{2\pi}{\frac{d^2K(\lambda)}{d\lambda^2}\Big|_{\hat{\lambda}}} \right\}^{1/2}, \tag{2.7}$$

em que, para cada y, $\hat{\lambda}$ satisfaz $dK(\hat{\lambda})/d\lambda = 0$ e $d^2K(\hat{\lambda})/d\lambda^2 < 0$ e, assim, maximiza $K(\lambda)$. Seja $\hat{\lambda}$ obtido pela solução de $K'(\hat{\lambda}) = y$ cuja nominação desta solução é *ponto de sela*. Pode-se imaginar que esta solução se refere ao estimador de máxima verossimilhança no

[1] ver Rudin (1971), *Princípios de Análise Matemática.*

modelo da família exponencial baseado na observação y. Expandindo $K(\lambda)-\lambda y$, referente ao expoente da equação (2.6), em série de Taylor, tem-se

$$K(\lambda)-\lambda y \approx K(\hat{\lambda})-\hat{\lambda} y+\frac{(\lambda-\hat{\lambda})^2}{2}K''(\hat{\lambda})$$

e associando esta aproximação com a equação (2.7), obtém-se

$$f(y) \approx \left\{\frac{1}{2\pi K''(\hat{\lambda})}\right\}^{1/2} \exp\{K(\hat{\lambda})-\hat{\lambda} y\}. \tag{2.8}$$

A equação (2.8) é denominada *expansão ponto de sela* para uma função densidade ou função de probabilidade e seu erro de aproximação é bem melhor que o da função obtida através da aproximação em série de Taylor (GOUTIS; CASELLA, 1999). A sua aplicabilidade depende apenas do conhecimento da f.g.c. $K(\lambda)$ e do cálculo de $\hat{\lambda}$ em $K'(\hat{\lambda})=y$. Esta equação é, em geral, não-linear.

Em muitas aplicações, a equação ponto de sela ($K'(\hat{\lambda})=y$) pode não ser resolvida analiticamente, mesmo quando a solução de $\hat{\lambda}$ existe. Mesmo assim, os métodos de ponto de sela podem ser aplicados resolvendo a equação numericamente. Usa-se o algoritmo de Newton-Raphson para calcular o ponto de sela, tendo este em geral bom desempenho desde que a função $K(\lambda)-\lambda y$, que é minimizada, seja convexa. Há ainda o método da secante para encontrar o ponto de sela. Esse método, geralmente, produz a resposta correta em uma iteração se $K'(\lambda)$ é linear, como é o caso do método Newton-Raphson. Como no algoritmo de Newton-Raphson, se $K''(\lambda)$ varia muito rapidamente, a convergência pode ser muito lenta, então, não existe a possibilidade de divergência para o método secante (KOLASSA, 1997).

Uma aplicação importante da expansão ponto de sela é a aproximação da função densidade ou função de probabilidade de uma soma ou média de variáveis aleatórias i.i.d.. Por exemplo, é possível determinar uma aproximação para a função densidade de uma soma ou média de variáveis aleatórias com distribuição exponencial de parâmetro α. Como a equação (2.8) é facilmente obtida através da f.g.c., tem-se que a função geradora de momentos da média amostral ($\bar{Y}=\sum_{i=1}^{n} Y_i/n$) é dada por $\phi_{\bar{Y}}(\lambda)=\phi(\lambda/n)^n$ e, portanto, a f.g.c. de $\bar{Y}$ é $K_{\bar{Y}}(\lambda)=nK(\lambda/n)$. Assim, uma aplicação direta de (2.8) produz a expansão ponto de sela da função densidade de $\bar{Y}$

$$f_{\bar{Y}}(\bar{y}) \approx \left\{\frac{n}{2\pi K''(\hat{\lambda})}\right\}^{1/2} \exp[n\{K(\hat{\lambda})-\hat{\lambda}\bar{y}\}]. \tag{2.9}$$

A qualidade da expansão ponto de sela pode ser, frequentemente, obtida pela multi-

plicação da aproximação da função densidade por uma constante de forma que sua integração resulte um. Uma versão da renormalização é exata para as funções densidades da normal, gama e normal inversa (DANIELS, 1980).

Exemplo 1. A aplicação da equação (2.8) é mostrada a seguir por meio da obtenção da função densidade da distribuição $\chi^2_{p,\alpha}$ não-central. Essa distribuição não apresenta forma fechada e sua função densidade é escrita como

$$f(y) = \sum_{k=0}^{\infty} \frac{y^{p/2+k-1}e^{-y/2}}{\Gamma(p/2+k)2^{p/2+k}} \frac{\alpha^k e^{-\alpha}}{k!},$$

em que p são os graus de liberdade e α é o parâmetro de não-centralidade. A função geradora de momentos (f.g.m.) é expressa em forma fechada por

$$\phi_Y(\lambda) = \frac{\exp\{2\alpha\lambda/(1-2\lambda)\}}{(1-2\lambda)^{p/2}}.$$

O ponto de sela da equação é obtido por $K'(\hat{\lambda}) = y$ de forma que

$$\hat{\lambda}(y) = \frac{-p+2y-\sqrt{p^2+8\alpha y}}{4y}$$

e, mediante a equação (2.8), é calculada a função densidade aproximada da distribuição $\chi^2_{p,\alpha}$ não-central por

$$f(y) \approx \left\{\frac{-4\pi(4\alpha+p-2p\lambda)}{(-1+2\lambda)^3}\right\}^{-1/2} \exp\left\{\frac{2\alpha\lambda}{1-2\lambda} - \frac{1}{2}p\log(1-2\lambda) + \frac{p}{4} - \frac{y}{2} + \frac{1}{4}\sqrt{p^2+8\alpha y}\right\}.$$

Exemplo 2. Tendo por base a equação (2.9), obtém-se a expansão ponto de sela para a média amostral de uma distribuição exponencial de parâmetro um. A f.g.c. desta distribuição é dada por $K(\lambda) = -\log(1-\lambda)$. O ponto de sela $\hat{\lambda}$ é obtido de $K'(\hat{\lambda}) = \bar{y}$. Dessa maneira, $\hat{\lambda}$ é obtido de $\hat{\lambda} = 1 - 1/\bar{y}$ e, tem-se ainda, que $K(\hat{\lambda}) = \log(\bar{y})$ e $K''(\hat{\lambda}) = \bar{y}^2$. Através da equação (2.9) a aproximação ponto de sela para a média amostral da distribuição exponencial com parâmetro um é

$$f_{\bar{Y}}(\bar{y}) \approx \left(\frac{n}{2\pi}\right)^{1/2} \bar{y}^{n-1} \exp\{-n(\bar{y}-1)\}.$$

Exemplo 3. A aplicação da equação (2.9) é agora introduzida para a média amostral da distribuição de Poisson. Sejam $Y_1,\ldots,Y_n$ variáveis aleatórias i.i.d. que seguem uma distribuição de Poisson com média α. A função geradora de cumulantes de Y_i é $K(\lambda) = \alpha\{\exp(\lambda)-1\}$ cujo ponto de sela iguala $\hat{\lambda} = \log(\bar{y}/\alpha)$. A equação (2.9) pode ser usada diretamente, sendo que a média $\bar{y} = s/n$ admite agora somente valores inteiros. Dessa forma, tem-se

$$\begin{aligned} f_{\bar{Y}}(\bar{y}) &\approx \left(\frac{n}{2\pi\bar{y}}\right)^{1/2} \exp\left[n\left\{\alpha\left(\frac{\bar{y}}{\alpha}-1\right)-\left(\log\frac{\bar{y}}{\alpha}\right)\bar{y}\right\}\right] \\ &= \left(\frac{n}{2\pi\bar{y}}\right)^{1/2} \frac{\alpha^{n\bar{y}}e^{n(\bar{y}-\alpha)}}{\bar{y}^{n\bar{y}+1/2}}. \end{aligned}$$

2.2.2 Expansão ponto de sela através da expansão de Edgeworth

Segundo Daniels (1987), um enfoque para o estudo da expansão de Edgeworth utiliza a idéia pioneira de Esscher (1932) e que foi explorada por Cramér (1938), Blackwell e Hodges (1959), Bahadur e Rao (1960), Barndorff-Nielsen e Cox (1979), Robinson (1982), entre outros.

Apesar da derivação original de ponto de sela de Daniels (1954) ter sido baseada na inversa da função característica, existe uma derivação alternativa que Reid (1988) denominou de "versão mais estatística" baseada na expansão de Edgeworth. A expansão de Edgeworth de uma distribuição é obtida expandindo a f.g.c. através da série de Taylor em torno de zero e invertendo-a em seguida.

Sejam $Y_1,\ldots,Y_n$ variáveis aleatórias i.i.d. de realizações de Y com função densidade $f(y)$, se Y for contínua ou função de probabilidade se Y for discreta. Define-se a família exponencial conjugada uniparamétrica, em que λ é o parâmetro canônico, por

$$f(y;\lambda) = \exp\{\lambda y - K(\lambda)\}f(y), \tag{2.10}$$

em que $K(t)$ é a f.g.c. da variável aleatória Y. A família exponencial conjugada (2.10) reproduz exatamente a função densidade $f(y)$ postulada para os dados quando $\lambda = 0$. O divisor necessário para normalizar a expressão $\exp(\lambda y)f(y)$ é igual à função geratriz de momentos (2.4). Pode-se, facilmente, verificar que a f.g.c. $K(t;\lambda)$ da família exponencial (2.10) é expressa em termos daquela $K(t)$ de Y por $K(t;\lambda) = K(t+\lambda) - K(\lambda)$.

Como exemplificação da equação (2.10), faz-se uso da distribuição gama, $G(\alpha,\beta)$, tal

que a sua função densidade conjugada é expressa como

$$f(y;\alpha,\beta,\lambda) = \frac{(\beta-\lambda)^{\alpha}}{\Gamma(\alpha)}\exp\{-(\beta-\lambda)y\}y^{\alpha-1},$$

em que a f.g.c. da função densidade conjugada é $K(t) = \alpha\log\left(\frac{\beta-\lambda}{\beta-\lambda-t}\right)$.

Um outro exemplo é ilustrado a partir da distribuição de Weibull com parâmetros $\alpha > 0$ e $\beta > 0$, sendo sua função densidade conjugada dada por

$$f(y;\alpha,\beta,\lambda) = \alpha^{\frac{\lambda}{\beta}+1}\beta y^{\beta-1}e^{-y(\alpha y^{\beta-1}-\lambda)}\left\{\Gamma\left(1+\frac{\lambda}{\beta}\right)\right\}^{-1}.$$

Definem-se $S = \sum_{i=1}^{n} Y_i$ e $S^* = (S-n\mu)/n^{1/2}\sigma$, em que μ e σ^2 são a média e variância de Y_i. Sejam $f_S(s;\lambda)$ e $K_S(t;\lambda)$ as funções densidade e geratriz de cumulantes de S relativas à família (2.10). Dessa forma, tem-se $K_S(t;\lambda) = nK(t+\lambda) - nK(\lambda)$ e, por inversão, vem

$$f_S(s;\lambda) = \exp\{s\lambda - nK(\lambda)\}f_S(s) \tag{2.11}$$

sendo $f_S(s) = f_S(s;0)$.

As funções densidades de S e S^* correspondentes à família (2.11) estão relacionadas por

$$f_S(s;\lambda) = f_{S^*}(y;\lambda)\frac{1}{\sqrt{nK''(\lambda)}}, \tag{2.12}$$

em que $y = \{s - nK'(\lambda)\}/\sqrt{nK''(\lambda)}$. Aproxima-se $f_{S^*}(y;\lambda)$ pela expansão de Edgeworth escolhendo convenientemente $y = 0$ para anular o termo de ordem $O(n^{-1/2})$. Esta escolha equivale a considerar a distribuição em (2.10) definida por $\hat{\lambda}$ satisfazendo a equação $K'(\hat{\lambda}) = s/n$. Pode-se interpretar $\hat{\lambda}$ como a estimativa de máxima verossimilhança (EMV) de λ baseada numa única observação s de (2.11).

A expansão de Edgeworth da soma padronizada é dada por

$$f_{S^*}(y) = \phi(y)\left\{1 + \frac{\rho_3}{6\sqrt{n}}H_3(y) + \frac{\rho_4}{24n}H_4(y) + \frac{\rho_3^2}{72n}H_6(y)\right\} + O(n^{-3/2}), \tag{2.13}$$

sendo $\phi(y)$ a f.d.p. da distribuição normal $N(0,1)$, $f_{S^*}(y)$ a função densidade da soma padronizada S^*, $H_r(y)$ o polinômio de Hermite de ordem r, $\rho_j = \rho_j(\lambda) = K^{(j)}(\lambda)/K''(\lambda)^{j/2}$ para $j = 3$ e 4, $K^{(j)}(\lambda) = d^jK(\lambda)/d\lambda^j$ e $\rho_3(\lambda)$ e $\rho_4(\lambda)$ sendo os cumulantes padronizados que medem a assimetria e a curtose da distribuição de Y. Aqui, $H_3(y) = y^3 - 3y$, $H_4(y) = y^4 - 6y^2 + 3$ e $H_6(y) = y^6 - 15y^4 + 45y^2 - 15$.

Logo, $f_S(s;\lambda) = f_{S^*}(0;\lambda)\frac{1}{\sqrt{nK''(\lambda)}}$. Agora, $f_{S^*}(0;\lambda)$ segue de (2.13), observando que

os cumulantes referentes a (2.10) são iguais a n vezes as derivadas de $K(\lambda)$

$$f_{S^*}(0;\hat{\lambda}) = \frac{1}{\sqrt{2\pi}}\{1+M(\hat{\lambda})+O(n^{-2})\}, \tag{2.14}$$

em que $M(\lambda)$ é um termo de ordem $O(n^{-1})$ dado por

$$M(\lambda) = \frac{3\rho_4(\lambda)-5\rho_3(\lambda)^2}{24n}.$$

Fazendo $\lambda = \hat{\lambda}$ em (2.11), explicitando $f_S(s)$ e usando (2.12) e (2.14), tem-se

$$f_S(s) = \frac{\exp\{nK(\hat{\lambda})-s\hat{\lambda}\}}{\sqrt{2n\pi K''(\hat{\lambda})}}\{1+M(\hat{\lambda})+O(n^{-2})\}. \tag{2.15}$$

A fórmula (2.15) para aproximar a função densidade de S é denominada *expansão ponto de sela da soma estocástica* e produz aproximações precisas para funções densidades baseadas nas suas funções geratrizes de cumulantes. O termo principal da equação (2.15) é chamado *expansão ponto de sela* para a função densidade da soma estocástica S proveniente de Y. Observe-se que o termo principal (2.15) só depende da função geratriz de cumulantes $K(\lambda)$. Esta fórmula é bem diferente da expansão de Edgeworth (2.13). Primeiro, para usar (2.15) é necessário calcular, além de $\hat{\lambda}$, a função geratriz de cumulantes de $K(\lambda)$ e não somente os seus 4 primeiros cumulantes. Entretanto, nas aplicações isso não apresenta grandes dificuldades.

O termo principal em (2.15) não é a função densidade da distribuição normal $N(0,1)$ e, embora seja sempre positivo, nem sempre a sua integral resulta em um. Assim, uma desvantagem de (2.15) é que nem sempre é fácil integrar o seu lado direito para obter uma aproximação para a função de distribuição de S. Entretanto, este termo pode ser normalizado. A expansão (2.15) é expressa em potências de ordem $O(n^{-1})$, enquanto a expansão de Edgeworth é dada em potências de ordem $O(n^{-1/2})$.

A expansão para a função densidade da média amostral $\bar{Y} = S/n$ segue diretamente de (2.15) como

$$f_{\bar{Y}}(y) = \left\{\frac{n}{2\pi K''(\hat{\lambda})}\right\}^{\frac{1}{2}} \exp[n\{K(\hat{\lambda})-\hat{\lambda}y\}]\{1+M(\hat{\lambda})+O(n^{-2})\}. \tag{2.16}$$

O termo principal em (2.16) é denominado *expansão ponto de sela para a função densidade (ou de probabilidade) da média amostral*.

Segundo Hinkley et al. (1990) e Cordeiro (1999), o interesse maior na inferência está

na obtenção de aproximações precisas para probabilidades do tipo $P(S \geq s)$ (ou $P(\bar{Y} \geq y)$) de uma amostra i.i.d. de n observações. Uma maneira de aproximar $P(S \geq s)$ é integrar numericamente a expansão ponto de sela representada pelo termo principal em (2.15), ou seja, calcular

$$P(S \leq s) = \int_{-\infty}^{s} \frac{\exp\{nK(\hat{\lambda}) - y\hat{\lambda}\}}{\sqrt{2n\pi K''(\hat{\lambda})}} dy.$$

A expansão de Edgeworth correspondente à função de distribuição de S^* é obtida de (2.13) por integração, sendo expressa por

$$F_{S^*}(y) = \Phi(y) - \phi(y)\left\{\frac{\rho_3}{6\sqrt{n}}H_2(y) + \frac{\rho_4}{24n}H_3(y) + \frac{\rho_3^2}{72n}H_5(y)\right\} + O(n^{-3/2}).$$

Aqui, $\Phi(\cdot)$ é a função de distribuição acumulada (f.d.a.) da distribuição normal $N(0,1)$ e os respectivos polinômios de Hermite de ordem r são: $H_2(y) = y^2 - 1$, $H_3(y) = y^3 - 3y$ e $H_5(y) = y^5 - 10y^3 + 15y$.

2.2.3 Expansão ponto de sela através de Lugannani-Rice

Uma forma alternativa simples de obter $P(S \leq s)$ até ordem $O(n^{-1})$, válida sobre todo o intervalo de variação de s, é devida a Lugannani e Rice (1980).

Integra-se a equação (2.15), em que $nK'(\hat{\lambda}) = s$, e, obtém-se, invertendo-a $\hat{\lambda} = \hat{\lambda}(x)$. Desde que $dx/d\hat{\lambda} = nK''(\hat{\lambda})$, vem

$$P(S \leq s) = \int_{-\infty}^{\hat{\lambda}(s)} \sqrt{\frac{nK''(\hat{\lambda})}{2\pi}} \exp[n\{K(\hat{\lambda}) - \hat{\lambda}K'(\hat{\lambda})\}] d\hat{\lambda}.$$

Considere ainda a seguinte mudança de variável

$$q = q(\hat{\lambda}) = \xi(\hat{\lambda})\sqrt{2\{\hat{\lambda}K'(\hat{\lambda}) - K(\hat{\lambda})\}},$$

tal que

$$\xi(\hat{\lambda}) = \begin{cases} -1, & \text{se } \hat{\lambda} < 0, \\ 1, & \text{se } \hat{\lambda} > 0. \end{cases}$$

A função $q(\hat{\lambda})$ é estritamente crescente e contínua. Com a mudança de variável $q = q(\hat{\lambda})$, a integral $P(S \leq s)$ é reescrita como

$$P(S \leq s) = \int_{-\infty}^{q_s} \sqrt{\frac{n}{2\pi K''(\hat{\lambda})}} \frac{q}{\hat{\lambda}} \exp(-nq^2/2) dq = \int_{-\infty}^{q_s} f(q)\phi(q; n^{-1}) dq,$$

em que $q_s = q(\hat{\lambda}(s))$, enquanto $\phi(u; n^{-1})$ indica a função densidade da distribuição normal $N(0, n^{-1})$. Além disso,

$$f(q) = \frac{q}{\hat{\lambda}\sqrt{K''(\hat{\lambda})}},$$

com $\hat{\lambda}$ expresso como uma função de q. Dessa forma, a integral $P(S \leq s)$ pode ser reescrita na forma que possibilita a aplicação da expansão de Laplace. Entretanto, esse método oferece três expressões distintas para aproximar de $P(S \leq s)$ dependendo de q_s ser negativo, igual a zero, ou positivo. Tem-se,

$$\begin{aligned} P(S \leq s) &= \int_{-\infty}^{q_s} \{1 + f(q) - 1\}\phi(q; n^{-1})dq \\ &= \Phi(\sqrt{n}q_s) + \int_{-\infty}^{q_s} \frac{f(q)-1}{q} q\phi(q; n^{-1})dq. \end{aligned}$$

Integrando por partes o último termo da soma, tomando $q\phi(q; n^{-1})$ como o fator diferencial, sendo ainda

$$\lim_{q \to -\infty} f(q)\phi(q; n^{-1})/q = 0,$$

tem-se

$$\begin{aligned} P(S \leq s) = \Phi(\sqrt{n}q_s) &- \frac{1}{n}\phi(q_s; n^{-1})\frac{f(q_s)-1}{q_s} \\ &+ \frac{1}{n}\int_{-\infty}^{q_s} \frac{d}{dq}\left(\frac{f(q)-1}{q}\right)\phi(q; n^{-1})dq, \end{aligned}$$

em que a última integral é ainda da forma $\int_{-\infty}^{q_s} f(q)\phi(q; n^{-1})dq$, com uma especificação diferente da função $f(\cdot)$. Essa integral é multiplicada por n^{-1}, de forma que sua contribuição é da ordem $O(n^{-1})$. Assim, tem-se a expansão assintótica

$$P(S \leq s) = \left\{\Phi(\sqrt{n}q_s) + \phi(q_s; n^{-1})\frac{1 - f(q_s)}{nq_s}\right\}\{1 + O(n^{-1})\}.$$

Escrevendo $\hat{r} = \sqrt{n}q_s$, obtém-se a aproximação conhecida como a expansão de Lugannani-Rice expressa como

$$P(S \leq s) = \Phi(\hat{r}) + \left(\frac{1}{\hat{r}} - \frac{1}{\hat{v}}\right)\phi(\hat{r}), \tag{2.17}$$

em que $\hat{r} = \xi(\hat{\lambda})[2n\{\hat{\lambda}K'(\hat{\lambda}) - K(\hat{\lambda})\}]^{1/2}$ e $\hat{v} = \hat{\lambda}\{nK''(\hat{\lambda})\}^{1/2}$.

As quantidades $\hat{r}$ e $\hat{v}$ podem ser interpretadas como a razão sinalizada de verossimilhanças e a estatística escore, respectivamente, para testar $\lambda = 0$ no modelo exponencial (2.11) definido para S.

A aproximação (2.17) é boa em quase todo intervalo de variação de s, exceto próximo ao ponto $s = E(S)$ ou $r = 0$, em que deve ser substituída pelo seu limite, quando $r \to 0$, dado por

$$P(S \leq s) = \frac{1}{2} + \frac{\hat{\rho}_3}{6\sqrt{2\pi n}}.$$

2.2.4 Expansão ponto de sela através da expansão de Daniels

Em Daniels (1954) é afirmado que é, freqüentemente, necessário aproximar as distribuições de algumas estatísticas cujas distribuições exatas não podem ser obtidas convenientemente. Segundo Reid (1988), a aproximação obtida por Daniels em 1954 é mais precisa que a aproximação de Edgeworth especialmente para n pequeno. Com o propósito de obter aproximações precisas para as distribuições de probabilidade da média amostral $\bar{Y}$ de n observações de uma distribuição com média $E(Y)$, Daniels (1987) propõe uma aproximação cujo erro relativo é controlado sobre todo o campo de variação de $\bar{Y}$.

Quando $\hat{\lambda} > 0$, a sua expansão de $P(S \geq s)$ até termos de ordem $O(n^{-1})$ é dada por

$$P(S \geq s) = \exp(n\hat{K} - s\hat{\lambda} + \hat{v}^2/2) \left[\{1 - \Phi(\hat{v})\} \left\{ 1 - \frac{\hat{\rho}_3 \hat{v}^3}{6\sqrt{n}} + \frac{1}{n} \left(\frac{\hat{\rho}_4 \hat{v}^4}{24} + \frac{\hat{\rho}_3^2 \hat{v}^6}{72} \right) \right\} \right.$$
$$\left. + \phi(\hat{v}) \left\{ \frac{\hat{\rho}_3(\hat{v}^2 - 1)}{6\sqrt{n}} - \frac{1}{n} \left(\frac{\hat{\rho}_4(\hat{v}^3 - \hat{v})}{24} + \frac{\hat{\rho}_3^2(\hat{v}^5 - \hat{v}^3 + 3\hat{v})}{72} \right) \right\} \right], \quad (2.18)$$

em que $\hat{\rho}_3 = \rho_3(\hat{\lambda})$, $\hat{\rho}_4 = \rho_4(\hat{\lambda})$, $\hat{K} = K(\hat{\lambda})$ e $\hat{v} = \hat{\lambda}\{nK''(\hat{\lambda})\}^{1/2}$. A aproximação de (2.18) com apenas os termos de ordem $O(\sqrt{n})$ fornece, em geral, bons resultados.

No caso em que $\hat{\lambda} < 0$, pode-se obter $P(S \geq s)$ até ordem $O(n^{-1/2})$ como

$$P(S \geq s) = H(-\hat{v}) + \exp(n\hat{K} - s\hat{\lambda} + \hat{v}^2/2) \times$$
$$\left[\{H(\hat{v}) - \Phi(\hat{v})\} \left(1 - \frac{\hat{\rho}_3 \hat{v}^3}{6\sqrt{n}} \right) + \phi(\hat{v}) \frac{\hat{\rho}_3(\hat{v}^2 - 1)}{6\sqrt{n}} \right],$$

em que $H(w) = 0,\ 1/2$ e 1 quando $w < 0$, $w = 0$, $w > 0$, respectivamente.

2.3 Identidades de Bartlett

Cordeiro (1999) afirma em seu livro que um problema natural que surge ao se usar os testes em grandes amostras é o de verificar se a aproximação de primeira ordem é adequada para a distribuição nula da estatística de teste em consideração. Para o caso

em que se trata de pequenas amostras, a aproximação de primeira ordem pode não ser satisfatória, conduzindo a taxas de rejeição bastante distorcidas. Na existência dessas situações, Bartlett (1937) propôs uma primeira idéia para melhorar os testes estatísticos. Ele considerou apenas a razão de verossimilhanças, computando o seu valor esperado segundo H até ordem $O(n^{-1})$, onde n é o tamanho da amostra.

Considera-se um conjunto de n observações independentes e identicamente distribuídas $y_1,\ldots,y_n$, que seguem qualquer distribuição regular uniparamétrica indexada por um parâmetro escalar desconhecido θ dada uma observação y. Seja $L(\theta;y)$ a verossimilhança total de um problema regular e, seja ainda, $\ell(\theta)=\log L(\theta;y)$ a log-verossimilhança. Assume-se que $\ell(\theta)$ satisfaz as condições de regularidade padrão e ainda $U_\theta = d\ell(\theta)/d\theta, U_{\theta\theta} = d^2\ell(\theta)/d\theta^2$, etc. No que se segue, usa-se a notação padrão para os cumulantes conjuntos de derivadas da log-verossimilhança:

$$\kappa_{\theta\theta}=E(U_{\theta\theta}),\ \kappa_{\theta\theta\theta}=E(U_{\theta\theta\theta}),\ \kappa_{\theta,\theta}=E(U_\theta^2)=-\kappa_{\theta\theta},$$
$$\kappa_{\theta,\theta\theta}=E(U_\theta U_{\theta\theta}),\ \kappa_{\theta\theta,\theta\theta}=E(U_{\theta\theta}^2)-\kappa_{\theta\theta}^2,\ \kappa_{\theta\theta\theta\theta}=E(U_{\theta\theta\theta\theta}),$$
$$\kappa_{\theta,\theta,\theta\theta}=E(U_\theta^2U_{\theta\theta})-\kappa_{\theta,\theta}\,\kappa_{\theta\theta},\ \kappa_{\theta,\theta,\theta,\theta}=E(U_\theta^4)-3\kappa_{\theta,\theta}^2 \quad \text{e} \quad \kappa_{\theta,\theta\theta\theta}=E(U_\theta U_{\theta\theta\theta}).$$

Denota-se, também, as derivadas dos cumulantes com sobrescritos como

$$\kappa_{\theta\theta}^{(\theta)}=d\kappa_{\theta\theta}/d\theta, \quad \kappa_{\theta\theta}^{(\theta\theta)}=d^2\kappa_{\theta\theta}/d\theta^2, \text{ etc.}$$

Outras identidades de Bartlett usuais são definidas como:

$$\kappa_{\theta,\theta}=-\kappa_{\theta\theta}, \quad \kappa_{\theta,\theta,\theta}=2\kappa_{\theta\theta\theta}-3\kappa_{\theta\theta}^{(\theta)}, \quad \kappa_{\theta,\theta\theta}=\kappa_{\theta\theta}^{(\theta)}-\kappa_{\theta\theta\theta},$$
$$\kappa_{\theta,\theta,\theta,\theta}=-3\kappa_{\theta\theta\theta\theta}+8\kappa_{\theta\theta\theta}^{(\theta)}-6\kappa_{\theta\theta}^{(\theta\theta)}+3\kappa_{\theta\theta,\theta\theta},$$
$$\kappa_{\theta,\theta,\theta\theta}=\kappa_{\theta\theta\theta\theta}-2\kappa_{\theta\theta\theta}^{(\theta)}+\kappa_{\theta\theta}^{(\theta\theta)}-\kappa_{\theta\theta,\theta\theta}, \quad \kappa_{\theta,\theta\theta\theta}=\kappa_{\theta\theta\theta}^{(\theta)}-\kappa_{\theta\theta\theta\theta}.$$

A grande vantagem das identidades de Bartlett é facilitar a obtenção dos cumulantes κ's, pois determinada parametrização pode conduzir a um cálculo direto simples de alguns cumulantes, sendo os demais calculados de forma indireta através destas identidades (CORDEIRO, 1999). Esses cumulantes têm como aplicabilidade principal o cálculo de viés de segunda ordem da EMV e no cálculo de correções de Bartlett e tipo-Bartlett para a estatística da razão de verossimilhanças, de maneira a melhorar os referidos testes, e tipo-Bartlett para a estatística escore. Para maiores detalhes sobre as Identidades de Bartlett e suas correções, sugere-se consultar Rodrigues (2006), Cordeiro (1999), Cordeiro e Ferrari (1991), Cysneiros e Cordeiro (2002), Cysneiros et al. (2009), entre outros.

2.3.1 Três Estatísticas Corrigidas

Os testes em grandes amostras, cujas distribuições de referência são qui-quadrado, mais conhecidos são: razão de verossimilhanças (w), escore (S) e Wald (W). Apresenta-se, aqui, três estatísticas corrigidas correspondentes aos testes em grandes amostras. Porém, na Seção 2.4, é apresentada a estatística proposta por Cordeiro e Ferrari (1998) cuja aplicabilidade será ilustrada para obtenção das probabilidades referentes às distribuições gama $G(\mu,\phi)$ e t-Student com ν graus de liberdade.

Para testar $H:\theta=\theta^{(0)}$ em qualquer distribuição regular uniparamétrica, em que $\theta^{(0)}$ é um valor especificado para θ, assume-se que $T\rightarrow\chi^2_1$ sob a hipótese nula H. Denote a log-verossimilhança total por $\ell_T(\theta)$ e a função escore total por $U_T(\theta)=d\ell_T(\theta)/d\theta$. Então, T pode tomar a forma de qualquer estatística a seguir

$$w=2[\ell_T(\widehat{\theta})-\ell_T(\theta^{(0)})],\quad S=U_T(\theta^{(0)})^2/(n\,\widetilde{\kappa}_{\theta,\theta}),$$

$$W=n(\widehat{\theta}-\theta^{(0)})^2\,\widehat{\kappa}_{\theta,\theta},\quad MW=n(\widehat{\theta}-\theta^{(0)})^2\,\widetilde{\kappa}_{\theta,\theta},$$

em que $\widehat{\kappa}_{\theta,\theta}$ e $\widetilde{\kappa}_{\theta,\theta}$ são a informação para uma observação avaliada em $\widehat{\theta}$ e $\theta^{(0)}$, respectivamente.

Tem-se que

$$P(T\leq x)=P(\chi^2_1\leq x)+O(n^{-1}).$$

e

$$P(T^*\leq x)=P(\chi^2_1\leq x)+O(n^{-3/2}).$$

A seguir, apresentam-se três propostas para corrigir a estatística T:

- Cordeiro e Ferrari (1991):

$$T^*=T\left[1-\frac{1}{n}\left(\alpha_1+\alpha_2T+\alpha_3T^2\right)\right].$$

- Kakizawa (1996):

$$K(T)=T^*+\frac{1}{4}[\alpha_3^2T+2\alpha_2\alpha_3T^2+2\alpha_1\alpha_3+\frac{4}{3}\alpha_2^2T^3 +3\alpha_1\alpha_2T^4+\frac{9}{5}\alpha_1^2T^5].$$

- Cordeiro et al. (1998):

$$\widetilde{T} = \sqrt{\frac{\pi}{3\alpha_1}} \exp\left(\frac{\alpha_2^2}{3\alpha_1} - \alpha_3\right) \times \left\{\Phi\left(\sqrt{6\alpha_1}\, T + \sqrt{\frac{2}{3\alpha_1}}\alpha_2\right) - \Phi\left(\sqrt{\frac{2}{3\alpha_1}}\alpha_2\right)\right\},$$

se $\alpha_1 > 0$ (neste caso, α_1 é sempre não negativo).

Quando $\alpha_1 = 0$ (e $\alpha_2 \neq 0$)

$$\widetilde{T} = \frac{1}{2\alpha_2} \exp(-\alpha_3)\{1 - \exp(-2\alpha_2 T)\}.$$

Diferentes α_i's são apresentados correspondendo às estatísticas de razão de verossimilhanças w, estatística escore S, estatística de Wald W e estatística de Wald modificada MW. Tem-se os seguintes α_i's para a estatística da razão de verossimilhanças w:

$$\alpha_1 = \frac{5\kappa_{\theta\theta\theta}^2 + 24\kappa_{\theta\theta}^{(\theta)}(\kappa_{\theta\theta}^{(\theta)} - \kappa_{\theta\theta\theta})}{12\,\kappa_{\theta\theta}^3} - \frac{\kappa_{\theta\theta\theta\theta} + 4(\kappa_{\theta\theta}^{(\theta\theta)} - \kappa_{\theta\theta\theta}^{(\theta)})}{4\kappa_{\theta\theta}^2},$$

$$\alpha_2 = \alpha_3 = 0.$$

Para a estatística escore S, tem-se que

$$\alpha_1 = \frac{-\kappa_{\theta,\theta,\theta}^2}{36\,\kappa_{\theta\theta}^3},$$

$$\alpha_2 = \frac{10\kappa_{\theta,\theta,\theta}^2 + 3\,\kappa_{\theta\theta}\kappa_{\theta,\theta,\theta,\theta} - 9\kappa_{\theta\theta}^3}{36\,\kappa_{\theta\theta}^3},$$

$$\alpha_3 = \frac{-5\kappa_{\theta,\theta,\theta}^2 - 3\kappa_{\theta\theta}\kappa_{\theta,\theta,\theta,\theta} + 9\kappa_{\theta\theta}^3}{12\,\kappa_{\theta\theta}^3}.$$

Para a estatística de Wald W, tem-se que

$$\alpha_1 = \frac{-44\kappa_{\theta\theta\theta}^2 + 120\kappa_{\theta\theta\theta}\kappa_{\theta\theta}^{(\theta)} - 81(\kappa_{\theta\theta}^{(\theta)})^2 + 12\kappa_{\theta\theta}\kappa_{\theta,\theta,\theta\theta}}{12\,\kappa_{\theta\theta}^3} - \frac{3\kappa_{\theta\theta}\kappa_{\theta,\theta,\theta,\theta}}{12\,\kappa_{\theta\theta}^3}$$

$$\alpha_2 = \frac{-10{\kappa_{\theta\theta\theta}}^2 + 48(2\kappa_{\theta\theta\theta} - 3\kappa_{\theta\theta}^{(\theta)})^2}{72\kappa_{\theta\theta}^3} + \frac{6(\kappa_{\theta\theta}^{(\theta)} - \kappa_{\theta\theta\theta})(17\kappa_{\theta\theta\theta} - 45\kappa_{\theta\theta}^{(\theta)})}{72\kappa_{\theta\theta}^3} + \frac{3\kappa_{\theta\theta,\theta\theta} + 20\kappa_{\theta\theta\theta}^{(\theta)} - 11\kappa_{\theta\theta\theta\theta} - 12\kappa_{\theta\theta}^{(\theta\theta)}}{12\kappa_{\theta\theta}^2},$$

$$\alpha_3 = -\frac{\kappa_{\theta\theta\theta}^2}{36\,\kappa_{\theta\theta}^3}.$$

E para a estatística de Wald modificada MW, tem-se que

$$\alpha_1 = \frac{-44\kappa_{\theta\theta\theta}^2 + 120\kappa_{\theta\theta\theta}\kappa_{\theta\theta}^{(\theta)} - 81(\kappa_{\theta\theta}^{(\theta)})^2 + 12\kappa_{\theta\theta}\kappa_{\theta,\theta,\theta\theta}}{12\,\kappa_{\theta\theta}^3} - \frac{3\kappa_{\theta\theta}\kappa_{\theta,\theta,\theta,\theta}}{12\,\kappa_{\theta\theta}^3}$$

$$\alpha_2 = \frac{63\kappa_{\theta\theta\theta}\kappa_{\theta\theta}^{(\theta)} - 22\kappa_{\theta\theta\theta}^2 - 45(\kappa_{\theta\theta}^{(\theta)})^2}{18\,\kappa_{\theta\theta}^3} + \frac{4\kappa_{\theta\theta\theta\theta} - 4\kappa_{\theta\theta\theta}^{(\theta)} - 4\kappa_{\theta,\theta,\theta\theta} - 3\kappa_{\theta,\theta,\theta,\theta}}{12\,\kappa_{\theta\theta}^2}$$

$$\alpha_3 = -\frac{(3\kappa_{\theta\theta}^{(\theta)} - \kappa_{\theta\theta\theta})^2}{36\,\kappa_{\theta\theta}^3}.$$

Tome-se a distribuição Poisson como forma de ilustração do uso dessas estatísticas (1, 2 e 3 são índices que se referem a α_1, α_2 e α_3). Desse modo, observa-se que

$$\begin{aligned}
&LR1 = -\frac{1}{6\theta},\\
&S1 = \frac{1}{36\theta}, \quad S2 = -\frac{7}{36\theta}, \quad S3 = \frac{1}{6\theta},\\
&W1 = \frac{1}{6\theta}, \quad W2 = \frac{1}{18\theta}, \quad W3 = \frac{1}{9\theta},\\
&MW2 = -\frac{7}{36\theta}, \quad MW3 = \frac{1}{36\theta}\ .
\end{aligned}$$

Ao exemplificar para a distribuição em série logarítmica, tem-se que

$$\begin{aligned}
LR1 = {} & \frac{1}{12\theta\{\theta+\log(1-\theta)\}^3}\left\{2\theta^4+6\theta^3\log(1-\theta)+8\theta^2(\theta+1)\log^2(1-\theta)\right.\\
& \left.-3\theta\left(\theta^2-2\theta-2\right)\log^3(1-\theta)+2\left(\theta^2-\theta+1\right)\log^4(1-\theta)\right\},
\end{aligned}$$

$$S1 = -\frac{\{2\theta^2+3\log(1-\theta)\theta+(\theta+1)\log^2(1-\theta)\}^2}{36\theta\{\theta+\log(1-\theta)\}^3},$$

$$\begin{aligned}
S2 = {} & \frac{1}{36\theta\{\theta+\log(1-\theta)\}^3}\left\{22\theta^4+66\theta^3\log(1-\theta)+\theta^2(28\theta+73)\log^2(1-\theta)\right.\\
& \left.-3\theta\left(\theta^2-12\theta-12\right)\log^3(1-\theta)+\left(7\theta^2+8\theta+7\right)\log^4(1-\theta)\right\},
\end{aligned}$$

$$\begin{aligned}
S3 = {} & \frac{1}{36\theta\{\theta+\log(1-\theta)\}^3}\left\{22\theta^4+66\theta^3\log(1-\theta)+\theta^2(28\theta+73)\log^2(1-\theta)\right.\\
& \left.-3\theta\left(\theta^2-12\theta-12\right)\log^3(1-\theta)+\left(7\theta^2+8\theta+7\right)\log^4(1-\theta)\right\},
\end{aligned}$$

$$\begin{aligned}
W1 = -{} & \frac{1}{12\theta\{\theta+\log(1-\theta)\}^3}\left\{2\theta^4+6\theta^3\log(1-\theta)\right.\\
& +8\theta^2(\theta+1)\log^2(1-\theta)-3\theta\left(\theta^2-2\theta-2\right)\log^3(1-\theta)\\
& \left.+2\left(\theta^2-\theta+1\right)\log^4(1-\theta)\right\},
\end{aligned}$$

$$
\begin{aligned}
W2=&\frac{1}{36\theta\{\theta+\log(1-\theta)\}^3}\left\{22\theta^4+6\theta^3(12\theta-1)\log(1-\theta)\right.\\
&+4\theta^2\left(6\theta^2+43\theta-20\right)\log^2(1-\theta)+\left(70\theta^2-46\theta\right)\log^4(1-\theta)\\
&\left.-2\log^4(1-\theta)+3\theta\left(25\theta^2+18\theta-18\right)\log^3(1-\theta)\right\},\\
W3=&-\frac{\left\{3\theta^2\log(1-\theta)+2\theta^2+(4\theta-2)\log^2(1-\theta)\right\}^2}{36\theta\{\theta+\log(1-\theta)\}^3},\\
MW2=&\frac{1}{36\theta\{\theta+\log(1-\theta)\}^3}\left\{22\theta^4+6\theta^3(6\theta+5)\log(1-\theta)\right.\\
&+3\theta^2(25\theta-6)\log^3(1-\theta)\\
&+\theta^2\left(24\theta^2+46\theta+1\right)\log^2(1-\theta)\\
&\left.+\left(43\theta^2-28\theta+7\right)\log^4(1-\theta)\right\},\\
MW3=&-\frac{\{4\theta^2+3(\theta+1)\log(1-\theta)\theta+(5\theta-1)\log^2(1-\theta)\}^2}{36\theta\{\theta+\log(1-\theta)\}^3}.
\end{aligned}
$$

2.4 Correção de Bartlett generalizada

Seja S^* uma estatística unidimensional com f.d.a. $F_{S^*}(y)$ e função densidade $f_{S^*}(y)$. Considera-se que S^* é obtida de uma amostra Y de n observações independentes tendo funções densidades que dependem de um vetor de parâmetros desconhecidos θ.

A idéia por trás do procedimento de modificação de S^* é baseado na suposição de que a função distribuição $F_{S^*}(y)$ possa ser formalmente expandida como

$$
F_{S^*}(y) = F_Z(y) + \sum_{i=1}^{m}(-1)^i\eta_i\frac{D^iF_Z(y)}{i!}, \tag{2.19}
$$

em que $D^iF_Z(y) = d^iF_Z(y)/dy^i$, quando $m \to \infty$ (vide McCullagh (1987, equação 5.6)).

A f.d.a. da estatística S^*, $F_{S^*}(y)$, pode ainda ser expandida como

$$
F_{S^*}(y) = F_Z(y) + A_1(y) + A_2(y) + O(n^{-3/2}), \tag{2.20}
$$

em que $F_Z(y)$ é a f.d.a. de uma variável aleatória escalar Z de referência (não necessariamente normal) usada para aproximar a distribuição de S^*, $A_1(y)$ e $A_2(y)$ são termos de ordens $O(n^{-1/2})$ e $O(n^{-1})$, respectivamente, que dependem de algumas diferenças dos cumulantes $(\kappa_i - \kappa_{i0})$ de S^* e Z, sendo κ_i o i-ésimo cumulante de S^* e κ_{i0} o i-ésimo cumulante de Z.

A forma (2.20) da f.d.a. de S^* sugere o uso de uma estatística modificada definida por

$$CF = S^* - b_1(S^*) - b_2(S^*),$$

em que $b_r(S^*) = O_p(n^{-r/2})$, para $r = 1$ e 2, são termos estocásticos aditivos como funções da estatística S^*. Cordeiro e Ferrari (1998) deduziram que $b_1(y) = -A_1(y)/f_Z(y)$ e $b_2(y) = -A_2(y)/f_Z(y) + A_1(y)^2 f'_Z(y)/\{2f_Z(y)^3\}$, supondo que $f_Z(y)$ é não-nula no suporte de S^*.

Assim, a estatística modificada CF cuja função de distribuição é $F_Z(y)$ até termos de ordem $O(n^{-1})$ é expressa como

$$CF = S^* \left[1 + \frac{A_1(S^*)}{f_Z(S^*)S^*} + \frac{1}{S^*} \left\{ \frac{A_2(S^*)}{f_Z(S^*)} - \frac{A_1(S^*)^2 f'_Z(S^*)}{2f_Z(S^*)^3} \right\} \right]. \tag{2.21}$$

O método que obtém (2.21) é formalmente válido e provado somente para que a distribuição de S^* inclua o termo de ordem $O(n^{-1})$ da expansão de Edgeworth. Tem-se, $P(CF \leq x) = P(Z \leq x) + O(n^{-3/2})$. O termo multiplicativo de S^* (2.21) é um tipo de ajuste estocástico envolvendo as funções $A_1(S^*)$ e $A_2(S^*)$, de ordens $O(n^{-1/2})$ e $O(n^{-1})$, deduzidas da expansão (2.20), da densidade $f_Z(y)$ com sua primeira derivada $f'_Z(y)$ e da estatística S^*.

O fator multiplicativo estocástico em (2.21) pode ser escrito como $1 + b(S^*, \eta_i, F_Z^{(i)}(y))$, em que a notação enfatiza a dependência das derivadas da função de distribuição $F_Z(y)$, dos "momentos formais" η_i's e da estatística não modificada S^*. Esses momentos formais são definidos em termos de diferenças entre os cumulantes de S^* e Z. Cordeiro e Ferrari (1998) definem este fator de ajuste como *correção de Bartlett generalizada* tal que este é um resultado geral que pode ser utilizado em muitos testes importantes na estatística e econometria.

Suponha que existem $\mu = \mu(\theta)$ e $\sigma = \sigma(\theta)$ tais que a estatística padronizada seja dada por $S^* = (S - \mu)/\sigma$, em que S é uma estatística que depende dos parâmetros n e θ. A estatística S^* tem média zero e variância unitária de forma a convergir em distribuição para uma variável aleatória com distribuição normal padrão.

Combinando a equação da expansão da f.d.a. em série de Edgeworth com (2.20), pode-se observar imediatamente que $A_1(y) = -\rho_3 \phi(y) H_2(y)/(6\sqrt{n})$ e $A_2(y) = -\phi(y)\{\rho_4 H_3(y)/24 + \rho_3^2 H_5(y)/72\}/n$. Substituindo esses resultados na equação (2.21), obtém-se

$$CF = S^* - \frac{\rho_3}{6\sqrt{n}}(S^{*2} - 1) + \frac{1}{12n} \left\{ \frac{\rho_3^2(4S^{*2} - 7S^*)}{3} - \frac{\rho_4(S^{*3} - 3S^*)}{2} \right\}. \tag{2.22}$$

A equação (2.22) é a transformação clássica polinomial de Cornish-Fisher para obter

normalidade quando os quantis estocásticos de ordem $O_p(n^{-3/2})$ e menores são omitidos, isto é, $CF \sim N(0,1) + O_p(n^{-3/2})$.

Exemplo 4. Mediante o uso da equação (2.19), apresenta-se a aproximação da distribuição gama $G(\mu,\phi)$ de média μ e parâmetro de dispersão ϕ considerando que é a f.d.a. da estatística S^*, em função da distribuição exponencial Z com média α (representando a f.d.a. da variável aleatória Z). Seja $F_Z(y) = 1 - e^{-y/\alpha}, y > 0, \alpha > 0$. Inicialmente, tem-se que

$$D^i F_Z(y) = (-1)^{i-1} \frac{e^{-y/\alpha}}{\alpha^i}, y > 0, \alpha > 0. \tag{2.23}$$

A relação dos momentos em função dos cumulantes para $r = 1,2$ e 3 é $\mu_1 = \kappa_1$, $\mu_2 = \kappa_2$, $\mu_3 = \kappa_3$ e para $r \geq 4$ pode ser expressa como

$$\mu_r = \kappa_r + \sum_{j=2}^{r-2} \binom{r-1}{j-1} \kappa_j \mu_{r-j}.$$

Assim, para $r = 4,5,6$ e 7 tem-se as seguintes identidades: $\mu_4 = \kappa_4 + 3\kappa_2^2$, $\mu_5 = \kappa_5 + 10\kappa_2\kappa_3$, $\mu_6 = \kappa_6 + 15\kappa_2\kappa_4 + 10\kappa_3^2 + 15\kappa_2^3$ e $\mu_7 = \kappa_7 + 10\kappa_2\kappa_5 + 80\kappa_2^2\kappa_3 + 20\kappa_3\kappa_4$. Sejam ε's as diferenças entre os cumulantes de S^* e Z, tais que

$$\varepsilon_i = (i-1)!\alpha^i \left[\frac{1}{\phi^{i-1}} \left(\frac{\mu}{\alpha} \right)^i - 1 \right].$$

Tem-se que os $\eta_i = O(\alpha^i) \times [\varepsilon_i]$ representam momentos centrais em função dos cumulantes correspondentes, que são iguais às diferenças entre os cumulantes de S^* e Z. Seja $[\varepsilon]_i$ um fator que representa uma combinação dos ε_i's.

A aproximação da distribuição gama $G(\mu,\phi)$ em função da distribuição exponencial de média α decorre substituindo na equação (2.19) a equação (2.23) e η_i. Logo,

$$F_{S^*}(y) = F_Z(y) - e^{-y/\alpha} \sum_{i=1}^{m} \frac{[\varepsilon]_i}{i!}.$$

Supondo que $m = 7$, então a aproximação para $F_{S^*}(y)$, que representa a distribuição gama $G(\mu,\phi)$, será

$$\begin{aligned} F_{S^*}(y) = F_Z(y) - e^{-y/\alpha} \{ & \varepsilon_1 + \frac{1}{2!}\varepsilon_2 + \frac{2}{3!}\varepsilon_3 + \frac{3}{4!}(2\varepsilon_4 + \varepsilon_2^2) + \frac{4}{5!}(6\varepsilon_5 + 5\varepsilon_2\varepsilon_3) \\ & + \frac{5}{6!}[24\varepsilon_6 + 9(2\varepsilon_4 + \varepsilon_2^2) + 3\varepsilon_2^3 + 8\varepsilon_3^2] + \frac{1}{7!}[720\varepsilon_7 + 40\varepsilon_2(6\varepsilon_5 + 5\varepsilon_2\varepsilon_3) \\ & + 160\varepsilon_2^2\varepsilon_3 + 120\varepsilon_3(2\varepsilon_4 + \varepsilon_2^2)] \}, \end{aligned}$$

em que $\alpha = \mu$ e $\varepsilon_i = \phi^{-(i-1)} - 1$.

Tabela 2.1: Resultados exato e aproximado da distribuição gama $G(\mu,\phi)$ pela distribuição exponencial de média um, sendo $\phi = 1,5,10,20$ e 30 para diferentes valores de y.

$y=1$	ϕ	1	5	10	20	30
Exato		0,6321206	0,9932620	0,9999546	1	1
Aproximado		0,6321206	0,9915655	0,9960130	0,9983550	0,9991863
$y=3$	ϕ	1	5	10	20	30
Exato		0,9502130	0,9999997	1	1	1
Aproximado		0,9502130	0,9988585	0,9994604	0,9997774	0,9998899
$y=5$	ϕ	1	5	10	20	30
Exato		0,9932620	1	1	1	1
Aproximado		0,9932620	0,9998455	0,9999270	0,9999699	0,9999850
$y=10$	ϕ	1	5	10	20	30
Exato		0,9999546	1	1	1	1
Aproximado		0,9999546	0,9999990	0,9999995	0,9999998	0,9999999

Na Tabela 2.1, apresentam-se os resultados exatos e aproximados para diferentes valores de μ e ϕ para a aproximação da distribuição gama $G(\mu,\phi)$ em função da distribuição exponencial de parâmetro α, sendo $\alpha = \mu = 1$. Observa-se que a aproximação (2.19) é adequada somente quando $\phi > 1$. Entretanto, pode ser usada quando $\phi < 1$ desde que $y \geq 10$. A Tabela 2.2 apresenta os resultados exato e aproximado para $y = 10$ e $\phi < 1$.

Tabela 2.2: Resultados exato e aproximado da distribuição gama $G(\mu,\phi)$ pela distribuição exponencial de média um, sendo $\phi < 1$ para $y = 10$.

$y=10$	ϕ	0,2	0,4	0,6	0,8
Exato		0,8646647	0,9816844	0,9975212	0,9996645
Aproximado		0,8511665	0,9965158	0,9995440	0,9998807

Exemplo 5. Mediante o uso da equação (2.19), apresenta-se abaixo a aproximação da função de distribuição t-Student com ν graus de liberdade, considerando que esta é a f.d.a. da estatística S^*, em função da distribuição normal padrão Z. Tem-se,

$$P(T_\nu \leq t) = \Phi(t) - \phi(t)\left[\frac{t(t^2+1)}{4\nu} + \frac{t(3t^6-7t^4-5t^2-3)}{96\nu^2}\right] + O(\nu^{-3}). \tag{2.24}$$

A demonstração da equação (2.24) pode ser encontrada em Fisher (1925). Em Johnson et al. (1995b), também, apresenta-se esta equação, mas é importante a ressalva de que é preciso realizar uma correção na equação (28.15) contida na página 375 do livro, pois a segunda parte da fórmula está multiplicada pela função acumulada da distribuição normal, mas a forma correta seria a função densidade da normal padrão.

Finner et al. (2008) investigaram o comportamento assintótico da f.d.p. e da f.d.a. da distribuição t-Student com $\nu > 0$ graus de liberdade para ν tendendo a infinito quando o

argumento $t = t_\nu$ da f.d.p. (f.d.a.) depende de ν e tende a $\pm\infty$. O respectivo artigo volta-se para a análise de algumas propriedades assintóticas da cauda da distribuição t-Student comparada com a distribuição normal padrão.

Objetiva-se, portanto, utilizar o método apresentado no artigo de Cordeiro e Ferrari (1998) para obter uma melhor aproximação da distribuição t-Student.

Em princípio, define-se $\nu = n^{1/2}$, $\nu^2 = n$ e $\nu^3 = n^{3/2}$ em que n é o tamanho amostral. Tem-se, então,

$$P(T_\nu \leq t) = \Phi(t) - \phi(t)\left[\frac{t(t^2+1)}{4n^{1/2}} + \frac{t(3t^6-7t^4-5t^2-3)}{96n}\right] + O(n^{-3/2}). \tag{2.25}$$

Seja agora $F_S(t) = \Phi(t) + A_1(t) + A_2(t) + O(n^{-3/2})$. Observa-se a partir de (2.25) que

$$A_1(t) = -\phi(t)\Big[\frac{t(t^2+1)}{4n^{1/2}}\Big]$$

e

$$A_2(t) = -\phi(t)\Big[\frac{t(3t^6-7t^4-5t^2-3)}{96n}\Big].$$

Considerando como base a distribuição normal, através da equação (6) de Cordeiro e Ferrari (1998), segue-se que

$$t^* = t\left[1 - \frac{A_1(t)}{\phi(t)t} - \frac{1}{t}\Big\{-\frac{A_1(t)A_1'(t)}{\phi(t)^2} + \frac{A_2(t)}{\phi(t)} + \frac{A_1^2(t)\phi'(t)}{2\phi^3(t)}\Big\}\right]$$

Observa-se, então, a necessidade da derivada de $A_1(t)$, sendo esta

$$A_1'(t) = -\phi(t)\Big[\frac{3t^2+1}{4\sqrt{n}}\Big] - \phi'(t)\Big[\frac{t(t^2+1)}{4\sqrt{n}}\Big].$$

Tem-se que $\frac{d^n\phi(t)}{dt^n} = \phi(t)(-1)^n H_n(t)$, sendo assim, $\phi'(t) = -\phi(t)t$, pois $H_1(t) = t$. Logo, obtém-se $A_1'(t) = \frac{\phi(t)}{4\sqrt{n}}\Big(t^4 - 2t^2 - 1\Big)$. Portanto,

$$t^* = t + \frac{t(t^2+1)}{4n^{1/2}} - \frac{(t^3+t)(t^4-2t^2-1)}{16n} + \frac{t^3(t^2+1)^2}{32n} + \frac{t(3t^6-7t^4-5t^2-3)}{96n}$$

ou

$$t^* = t + \frac{t(t^2+1)}{4n^{1/2}} + \frac{t(5t^2+1)(t^2+3)}{96n}. \tag{2.26}$$

Pode-se, assim, expressar a equação (2.26) como

$$t^* = t + \frac{Q_1(t)}{\sqrt{n}} + \frac{Q_2(t)}{n},$$

em que $Q_1(t) = \frac{t(t^2+1)}{4}$ e $Q_2(t) = \frac{t(5t^2+1)(t^2+3)}{96}$.

Seguindo Cordeiro e Ferrari (1998, p. 515), deseja-se obter $F_{T_\nu}(t^*) = \Phi(t)$, em que t^* é expresso pela equação (2.26). Dessa forma, é necessário obter uma equação que expresse t em termos de t^*. Esta equação é deduzida a partir da equação de inversão (6.74) em Kendall (1945, p. 167), sendo esta expressa por

$$\begin{aligned} t - t^* &= g(t) = g(t^* + t - t^*) \\ &= g(t^*) + g'(t^*)g(t^*) + g(t^*)g'^2(t^*) + \frac{1}{2}g(t^*)^2 g''(t^*) + g(t^*)g'^3(t^*) + \frac{3}{2}g(t^*)^2 g'(t^*)g''(t^*) \\ &\quad + \frac{1}{6}g(t^*)^3 g'''(t^*) + \cdots \end{aligned}$$

Da equação (2.26), tem-se, então, $t - t^* = g(t)$, em que $g(t) = -\frac{Q_1(t)}{\sqrt{n}} - \frac{Q_2(t)}{n}$. Para utilizar a equação dada por Kendall (1945), torna-se necessário obter as derivadas de $g(t)$. Portanto,

$$\begin{aligned} g'(t) &= -\frac{3t^2+1}{4\sqrt{n}} - \frac{25t^4+48t^2+3}{96n}, \\ g''(t) &= -\frac{3t}{2\sqrt{n}} - \frac{t(25t^2+24)}{24n} \quad \text{e} \\ g'''(t) &= -\frac{3}{2\sqrt{n}} - \frac{25t^2+8}{8n}. \end{aligned}$$

Após determinar as derivadas de $g(t)$, obtém-se a expansão de t em função de t^*, ou seja,

$$\begin{aligned} t \approx t^* \Big\{ & 1 - \frac{(t^{*2}-1)}{4\sqrt{n}} + \frac{(13t^{*4}+8t^{*2}+3)}{96n} + \frac{t^{*2}(2t^{*4}-1)}{24n^{3/2}} \\ & + \frac{455t^{*8} - 656t^{*6} - 1158t^{*4} - 408t^{*2} - 9}{9216n^2} \\ & + \frac{8970t^{*10} + 30745t^{*8} + 31464t^{*6} + 12810t^{*4} + 1950t^{*2} + 45}{36864n^{5/2}} \\ & + \frac{656760t^{*10} + 1174855t^{*8} + 827808t^{*6} + 228375t^{*4} + 19080t^{*2} + 297}{884736n^3} \\ & + \frac{25500t^{*14} + 185500t^{*12} + 479115t^{*10} + 528055t^{*8} + 239250t^{*6} + 40194t^{*4} + 2295t^{*2} + 27}{884736n^{7/2}} \\ & + \frac{178125t^{*16} + 1632000t^{*14} + 5586000t^{*12} + 8765120t^{*10} + 6213130t^{*8} + 1733760t^{*6}}{84934656n^4} \\ & + \frac{201096t^{*4} + 8640t^{*2} + 81}{84934656n^4} \Big\}. \end{aligned}$$

Tabela 2.3: Valores aproximados da estatística t para diferentes valores de $\nu = \sqrt{n}$.

ν	t	$Exato$	$\Phi(t)$	(2.25)	(2.26)	(2.27)
	0,0	0,0	0,5	0,5	0,5	0,5
2	0,5	0,6666667	0,691462	0,666085	0,659960	0,666608
	1,5	0,8638034	0,933193	0,862118	0,738720	1,000000
	2,0	0,9082483	0,977250	0,893733	0,493767	1,000000
	0,0	0,0	0,5	0,5	0,5	0,5
3	0,5	0,674276	0,691462	0,674071	0,671365	0,674171
	1,5	0,884708	0,933193	0,884065	0,836974	0,961253
	2,0	0,930337	0,977250	0,925134	0,795708	1,000000
	0,0	0,0	0,5	0,5	0,5	0,5
4	0,5	0,678170	0,691462	0,678242	0,676724	0,678279
	1,5	0,896404	0,933193	0,895693	0,871289	0,915385
	2,0	0,940904	0,977250	0,939498	0,881713	0,999900
	0,0	0,0	0,5	0,5	0,5	0,5
5	0,5	0,6808506	0,691462	0,680801	0,679831	0,680819
	1,5	0,9030482	0,933193	0,902879	0,888018	0,908915
	2,0	0,9490303	0,977250	0,947690	0,915821	0,987824
	0,0	0,0	0,5	0,5	0,5	0,5
6	0,5	0,68256	0,691462	0,682530	0,681858	0,682541
	1,5	0,9078596	0,933193	0,907757	0,897773	0,910083
	2,0	0,9537868	0,977250	0,952973	0,932968	0,971901
	0,0	0,0	0,5	0,5	0,5	0,5
7	0,5	0,6837964	0,691462	0,683778	0,683284	0,683784
	1,5	0,9113508	0,933193	0,911284	0,904119	0,912351
	2,0	0,9571903	0,977250	0,956659	0,942987	0,965457
	0,0	0,0	0,5	0,5	0,5	0,5
8	0,5	0,684732	0,691462	0,684719	0,684341	0,684724
	1,5	0,9139984	0,933193	0,913952	0,908563	0,914511
	2,0	0,9597419	0,977250	0,959376	0,949461	0,963790
	0,0	0,0	0,5	0,5	0,5	0,5
9	0,5	0,6854644	0,691462	0,685455	0,685157	0,685458
	1,5	0,9160747	0,933193	0,916042	0,911841	0,916365
	2,0	0,9617236	0,977250	0,961461	0,953949	0,963869
	0,0	0,0	0,5	0,5	0,5	0,5
10	0,5	0,6860532	0,691462	0,686047	0,685805	0,686049
	1,5	0,9177463	0,933193	0,917722	0,914356	0,917924
	2,0	0,9633060	0,977250	0,963111	0,957227	0,964526
	0,0	0,0	0,5	0,5	0,5	0,5
20	0,5	0,6887341	0,691462	0,688733	0,688673	0,688734
	1,5	0,9253821	0,933193	0,925379	0,924577	0,925393
	2,0	0,9703672	0,977250	0,970341	0,969073	0,970408
	0,0	0,0	0,5	0,5	0,5	0,5
30	0,5	0,6896385	0,691462	0,689638	0,689611	0,689638
	1,5	0,9279670	0,933193	0,927966	0,927615	0,927970
	2,0	0,9726875	0,977250	0,972679	0,972143	0,972695
	0,0	0,0	0,5	0,5	0,5	0,5
100	0,5	0,6909132	0,691462	0,690913	0,690911	0,690913
	1,5	0,9316175	0,933193	0,931617	0,931587	0,931618
	2,0	0,9758940	0,977250	0,975894	0,975849	0,975894
	0,0	0,0	0,5	0,5	0,5	0,5
150	0,5	0,6910961	0,691462	0,691096	0,691095	0,691096
	1,5	0,9321419	0,933193	0,932142	0,932128	0,932142
	2,0	0,9763472	0,977250	0,976347	0,976327	0,976347

E, considerando termos até ordem $O(n^{-1})$, tem-se

$$t \approx t^* - \frac{t^*(t^{*2}+1)}{4\sqrt{n}} - \frac{t^*(5t^{*2}+1)(t^{*2}+3)}{96n} + \frac{t^*(t^{*2}+1)(3t^{*2}+1)}{16n}$$
$$= t^* - \frac{t^*(t^{*2}+1)}{4\sqrt{n}} + \frac{t^*(13t^{*4}+8t^{*2}+3)}{96n}. \quad (2.27)$$

Para diferentes valores de ν, apresenta-se na Tabela 2.3 a probabilidade exata para a distribuição t-Student, a probabilidade aproximada baseada na distribuição normal padrão (primeiro termo do lado direito de (2.25)), a probabilidade aproximada baseada na equação (2.25) e a probabilidade da normal padrão utilizando o valor da estatística t definida pela equação (2.27), isto é, a f.d.a. $\Phi(\cdot)$ no valor de t correspondente à equação (2.27).

Na Tabela 2.3 observa-se ainda que, à medida que ν aumenta os valores obtidos para (2.25), pela aproximação de t obtida através das equações (2.26) e (2.27) sob a distribuição normal padrão, se aproximam dos valores exatos da distribuição t-Student. Para o caso referente a f.d.a. $\Phi(\cdot)$ no valor de t correspondente à equação (2.26), observa-se que os valores das probabilidades aproximam-se do valor exato apenas à medida que se aumentam os graus de liberdade (g.l.). Para o caso referente a f.d.a. $\Phi(\cdot)$ no valor de t correspondente à equação (2.27), nota-se que os correspondentes valores das probabilidades se aproximam do valor exato da distribuição t-Student para valores pequenos de g.l. e para valores de t pequenos (exceto quando $\nu < 5$) e, ainda, à medida que o grau de liberdade é aumentado esse valor de probabilidade se torna ainda mais próximo do valor exato para qualquer valor de t apresentado.

Dessa maneira, baseado em Cordeiro e Ferrari (1998), vê-se a possibilidade de uso da propriedade $F_{T_\nu}(t^*) = \Phi(t)$, em que t é agora expresso pela equação (2.27), ou seja, fazer uso da equação (2.27) em termos da distribuição normal padrão para obter valores de probabilidade de uma distribuição t-Student com ν graus de liberdade.

2.5 Considerações Finais

Neste Capítulo foram apresentadas algumas expansões assintóticas utilizadas para obter a expansão ponto de sela. Como foi mencionado anteriormente, as expansões ponto de sela apresentam diversas aplicações na estatística, a exemplo da grande precisão em aproximar funções densidade e de distribuição da soma e da média de variáveis aleatórias i.i.d.. Ainda, mediante aproximação da função de distribuição proposta por Cordeiro e Ferrari (1998), foi possível obter uma aproximação da distribuição gama $G(\mu, \phi)$ em fun-

ção da distribuição exponencial de média α. E, ainda, ao utilizar a estatística, também, proposta por Cordeiro e Ferrari (1998), foi obtida uma estatística que ao ser usada em função de $\Phi(t)$, resulta no valor aproximado da probabilidade da distribuição t-Student com ν graus de liberdade. Isto, portanto, corresponde à propriedade $F_{T_\nu}(t^*) = \Phi(t)$. É importante observar que melhores resultados dessa relação entre a estatística (2.27) e a f.d.a. da distribuição normal são obtidos quando $\nu > 9$.

3 As Distribuições Beta Generalizadas

3.1 A Distribuição Beta

A distribuição beta é uma das mais usadas para modelar experimentos aleatórios que produzem resultados no intervalo $(0,1)$, dada a grande flexibilidade de ajuste de seus parâmetros. Esse fato a torna a mais flexível da família de distribuições. Ela tem relação com várias das mais conhecidas distribuições univariadas. As distribuições beta são muito versáteis e podem modelar uma grande variedade de incertezas. Muitas das distribuições finitas encontradas na prática podem ser facilmente transformadas na distribuição beta padronizada. Bury (1999) lista um conjunto de aplicações da distribuição beta em engenharia. Janardan e Padmanabhan (1986) modelam variáveis hidrológicas usando a distribuição beta. McNally (1990) utiliza a distribuição beta no estudo de algumas variáveis que afetam a reprodutibilidade de vacas. Graham e Hollands (1990) e Milyutin e Yaromenko (1991) usam a distribuição beta para estudar índices relacionados à transmissão da radiação solar. A potência de sinais de radar é modelada por Maffet e Wackerman (1991) através da distribuição beta. Wiley et al. (1989) desenvolvem um modelo beta para estimar a probabilidade de transmissão de HIV durante o contato sexual entre um indivíduo infectado e um indivíduo sadio. Johnson et al. (1995b, p. 235) observam: “the beta distributions are among the most frequently employed to model theorical distributions.”

Diz-se que uma variável aleatória Y tem distribuição beta padrão se sua função densidade de probabilidade (f.d.p.) é definida como

$$f(y;a,b) = \frac{1}{B(a,b)} y^{a-1}(1-y)^{b-1}, \quad a>0,\ b>0, \quad y \in (0,1), \tag{3.1}$$

em que $B(a,b) = \Gamma(a)\Gamma(b)/\Gamma(a+b)$ é a função beta completa, sendo $\Gamma(\cdot)$ a função gama, isto é, $\Gamma(p) = \int_0^\infty w^{p-1}\exp(-w)\mathrm{d}w$ com $p>0$.

A densidade (3.1) é unimodal para $a>0,\ b>1$ e não é unimodal quando $a<1,\ b<1$

ou $(a-1)(b-1) \leq 0$. Além disso, a distribuição beta, representada aqui por $beta(a,b)$, tem a propriedade de quase-simetria, isto é, se $Y \sim beta(a,b)$, então $1-Y \sim beta(a,b)$. Os casos especiais mais importantes da distribuição beta são: a uniforme $(a=b=1)$ e a distribuição potência $(a>0, b=1)$. Outro caso especial ocorre quando $a=b=1/2$, obtendo-se a distribuição arcosseno, e quando $b=1-a$, $0<a<1$, tem-se a distribuição arcosseno generalizada. As distribuições uniforme e arcosseno são distribuições simétricas em relação às suas respectivas médias.

Uma transformação linear da forma $Z=t+hY$, $-\infty<t<\infty$, $h>0$, fornece uma forma geral da distribuição beta com função densidade da forma

$$f(z;a,b)=\frac{1}{B(a,b)}(z-t)^{a-1}(h+t-z)^{b-1}, \quad t<z<t+h,$$

ou seja, $Z \sim beta(a,b,t,h)$. A propriedade de quase-simetria se verifica para a distribuição $beta(a,b,t,h)$, que também é simétrica em torno da sua média, sendo esta dada por $t+h/2$. O caso mais comum da distribuição beta, portanto, é dado quando $t=0$ e $h=1$, correspondendo a equação (3.1).

A função de distribuição acumulada, $F_Y(y)$, da distribuição beta é definida como

$$F_Y(y)=\frac{B_y(a,b)}{B(a,b)},$$

em que $B_y(a,b)=\int_0^y z^{a-1}(1-z)^{b-1}\mathrm{d}z$ é a função beta incompleta. O r-ésimo momento da distribuição $beta(a,b)$ pode ser obtido através da expressão

$$E(Y^r)=\mu_r=\frac{B(r+a,b)}{B(a,b)}, \quad r=0,1,2,\ldots.$$

Dessa forma, temos que $E(Y)=p/(p+q)$, $E(Y^2)=p(p+1)/[(p+q)(p+q+1)]$, e conseqüentemente, $\mathrm{Var}(Y)=pq/[(p+q)^2(p+q+1)]$.

3.1.1 Aplicação da Distribuição Beta em Estudos Agrários

Diversos são os estudos baseados na análise de variáveis definidas no intervalo $(0,1)$, tais como porcentagens ou proporções, tendo estas o modelo de regressão beta como um dos mais utilizados. Com o intuito de ilustrar a aplicabilidade da distribuição beta em estudos agrários, apresenta-se o modelo de regressão beta proposto por Ferrari e Cribari-Neto (2004), o qual sugere uma reparametrização da distribuição beta considerando a média da resposta e um parâmetro de dispersão. Sejam $\alpha=E(Y)=p/(p+q)$ e $\phi=p+q$, isto é, $p=\alpha\phi$ e $q=(1-\alpha)\phi$. Assim, a densidade de y, em (3.1), pode ser expressa como

$$f(y;\alpha,\phi) = \frac{\Gamma(\phi)}{\Gamma(\alpha\phi)\Gamma((1-\alpha)\phi)} y^{\alpha\phi-1}(1-y)^{(1-\alpha)\phi-1}, \quad 0<y<1,$$

em que $0<\alpha<1$ e $\phi>0$. Mediante isto, a média e a variância passam a ser expressadas como

$$E(y) = \alpha \quad e \quad Var(y) = \frac{V(\alpha)}{1+\phi},$$

em que $V(\alpha) = \alpha(1-\alpha)$ é a função de variância, α é a média da variável resposta e ϕ é interpretado como um parâmetro de precisão (dispersão) no sentido que, para α fixo, quanto maior for o valor de ϕ, menor a variância de y (para maiores detalhes ver Miyashiro (2008) e Ferrari e Cribari-Neto (2004)).

Realiza-se uma análise tendo como modelo base o modelo de regressão beta com a finalidade de fazer um ajuste referente a produção total de leite das vacas da raça SINDI no período de 1987 a 1997, sendo esta representada pela variável y, com respeito a duração da lactação (x_{t1}) e, também, com relação ao mês do parto (x_{t2}). Este conjunto de dados contém 107 observações sendo estas correspondentes apenas ao primeiro parto para as 107 lactações (ver Tabela 3.1). As vacas são de propriedade da fazenda Carnaúba, pertencente à AMDA (Agropecuária Manoel Danta Ltda), situada no município de Taperoá, microrregião do Cariri Ocidental do Estado da Paraíba. Como a variável resposta não está restrita ao intervalo unitário $(0,1)$, faz-se uma padronização da variável de forma a modelar $(y_t - min)/(max - min)$ ao invés de modelar y diretamente, em que min e max referem-se ao valor mínimo e máximo de y, respectivamente, e $t = 1,\ldots,107$. Seja, ainda, a variável explicativa relativa ao mês do parto (x_{t2}), fez-se uma modificação nesta com o intuito de produzir uma variável *dummy* e, assim, torná-la uma variável semestral a qual foi definida como $DSEM$.

Considere-se que as observações $y_1,\ldots,y_n$ são independentes e seguem a distribuição beta com média α_t, $t = 1,\ldots,n$, e parâmetro de precisão ϕ desconhecido. O modelo a ser ajustado é, então, expresso por

$$g(\alpha_t) = \delta_0 + \delta_1 x_{t1} + \ldots + \delta_k x_{tk}. \tag{3.2}$$

Ajusta-se o modelo de regressão beta apresentado em (3.2) utilizando a função de ligação *logit*. Os resultados das estimativas para o modelo ajustado estão especificadas na Tabela 3.2, onde observa-se que a variável *dummy* ($DSEM$) é não significativa para o modelo de regressão. Por conseguinte, faz-se um novo ajuste sem esta variável de

Tabela 3.1: Conjunto de dados referente às vacas da raça SINDI.

N	$PLTOTAL(y_t)$	$DULAC(x_{t1})$	$MESPARTO(x_{t2})$	N	$PLTOTAL(y_t)$	$DULAC(x_{t1})$	$MESPARTO(x_{t2})$
1	1800,8	295	4	55	2133,3	297	2
2	3117	343	4	56	1916,2	291	2
3	1201,9	235	4	57	1734,4	267	2
4	699,6	141	6	58	2075	284	12
5	2094,2	304	2	59	1663,9	279	4
6	1802,4	274	12	60	2177,6	268	2
7	1937,3	238	5	61	1638,5	271	4
8	2147,7	276	4	62	960,3	101	12
9	2511,5	309	2	63	1912,8	282	5
10	1675,4	249	8	64	681	69	2
11	3031	289	12	65	2726,5	321	4
12	1143,4	236	5	66	2219,4	286	6
13	1516,9	261	2	67	2564,4	297	2
14	1725,1	297	3	68	2625,1	293	4
15	1706,5	298	3	69	1471,1	143	5
16	2270,4	283	2	70	1860,1	218	3
17	1769,5	321	3	71	836,6	86	4
18	1987	321	3	72	1433,7	155	1
19	2291,7	288	10	73	2176,9	218	1
20	1203,4	231	12	74	2059,2	193	1
21	1736,9	322	4	75	1822,6	187	1
22	1508,1	265	9	76	1893	177	1
23	2261,9	297	12	77	1898,8	175	1
24	1849,8	297	12	78	1846	211	12
25	2023,8	317	12	79	1572,2	167	12
26	1856,9	302	1	80	2433,5	317	4
27	2928	298	1	81	1276,1	163	6
28	2498,7	299	1	82	2107,4	286	3
29	1790,8	276	9	83	2664	328	4
30	2094,1	267	11	84	2244,3	291	4
31	2154,9	304	12	85	1186,2	122	10
32	2340	314	3	86	2825,8	315	3
33	2031,8	312	1	87	2672,6	306	6
34	2305,4	324	3	88	1446	149	5
35	954,3	237	5	89	2034,8	307	5
36	1930,4	292	7	90	2426,8	318	5
37	2523,4	294	10	91	1874,4	297	6
38	2322,1	328	2	92	2539,6	285	7
39	2211,7	315	3	93	1583	141	1
40	1659,5	308	3	94	1298,7	184	7
41	2123,1	309	3	95	2773,8	256	6
42	1832,1	317	2	96	1318,5	178	12
43	1580,7	311	3	97	2346,4	268	10
44	2353,6	326	2	98	940,4	111	3
45	754,2	146	5	99	1490,2	170	10
46	1617,7	254	9	100	2517,1	306	9
47	2251,6	292	12	101	2791	301	3
48	2544,4	283	12	102	1514,8	155	8
49	2558,3	306	1	103	549,6	80	3
50	2553,6	292	12	104	847,5	105	1
51	2134	298	12	105	730,8	97	10
52	2200,8	302	12	106	693,4	77	1
53	1863,5	287	3	107	1450	165	1
54	1537	260	2	-	-	-	-

Nota: PLTOTAL, DULAC e MESPARTO correspondem, respectivamente, a produção total de leite, duração da lactação e ao mês em que ocorreu o parto.

forma a obter um novo modelo. As estimativas desse novo modelo são apresentadas na Tabela 3.3. A Figura 3.1 corresponde a análise dos resíduos do modelo ajustado sem a variável *dummy*. Observa-se a indicação de alguns poucos pontos residuais e um ponto, a princípio, influente indicado pela observação 103. Dessa forma, faz-se a remoção da provável observação influente e realiza-se uma nova estimativa dos parâmetros do modelo ajustado.

Tabela 3.2: Estimativas dos parâmetros referentes ao modelo de regressão beta.

Parâmetro	Estimativa	Erro Padrão	p-valor
Constante	-2,62786	0,02235	0,0000
DULAC	0,00962	8,13 $\times 10^{-4}$	0,0000
DSEM1	0,10942	0,01109	0,3240
ϕ	14,38607	1,910381	-

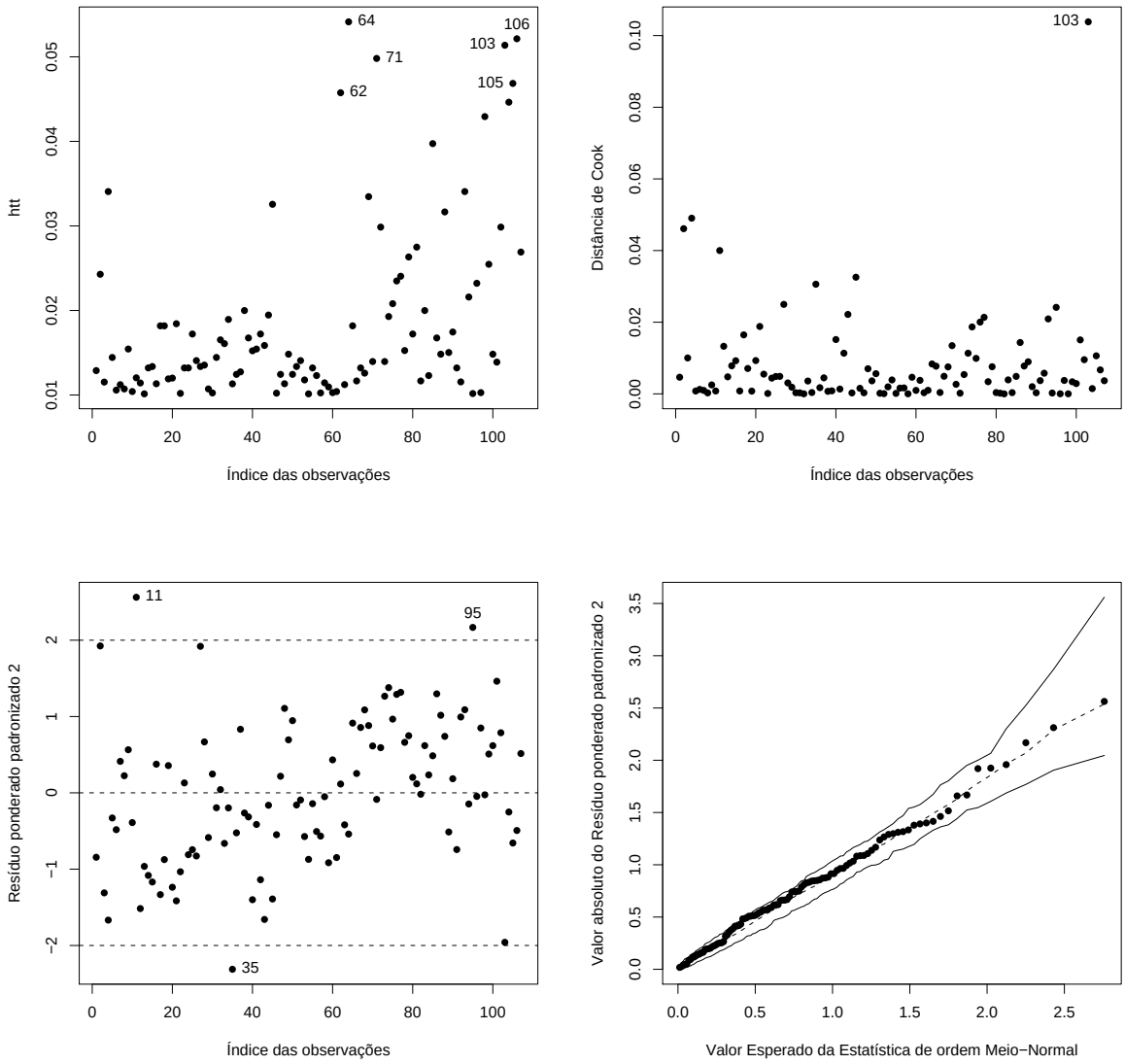

Figura 3.1: Análise dos resíduos do modelo de regressão beta ajustado.

Utiliza-se uma medida de variação percentual ($VP_{\hat{\theta}}$) com o intuito de avaliar a influência que as observações com características distintas das demais (destacadas como alavanca, aberrante e/ou influente) têm sobre as estimativas dos parâmetros de regressão e dispersão. Essa medida é expressa como

$$VP_{\hat{\theta}} = \frac{\hat{\theta}_{-ponto(s)} - \hat{\theta}}{\hat{\theta}},$$

em que $\hat{\theta}_{-ponto(s)}$ é a estimativa do parâmetro θ sem o(s) ponto(s) com característica(s) distinta(s) dos demais e $\hat{\theta}$ é a estimativa de θ com todos os pontos no modelo. Pode ser visto, portanto, na Tabela 3.4, a variação percentual para os parâmetros do modelo ajustado quando retirada a observação 103 que é aparentemente um ponto influente no modelo. Mediante isto, observa-se que a variação percentual dos parâmetros é pequena,

porém, acima da taxa de 2.8% indicando que a observação 103 muda as conclusões inferenciais, isto é, causa uma pequena influência sob a variável $DULAC$.

Tabela 3.3: Estimativas dos parâmetros referentes ao modelo de regressão beta sem a variável *dummy*.

Parâmetro	Estimativa	Erro Padrão	p-valor
Constante	-2,5811614	0,02184	0,0000
DULAC	0,0095756	8,13 $\times 10^{-4}$	0,0000
ϕ	14,25315	1,892158	-

Tabela 3.4: Variação percentual das estimativas dos parâmetros do modelo de regressão beta ajustado, sem a observação 103.

	Todas obs	Sem obs 103	
Parâmetro	Estimativa	Estimativa	$VP_{\hat{\theta}}$
Constante	-2,5811614	-2,4796470	-3,93%
DULAC	0,0095756	0,0092192	-3,72%
ϕ	14,25315	14,61016	2,50%

Dessa forma, sendo recomendável a remoção da observação 103, o modelo final obtido segue a seguinte estrutura

$$g(\alpha_t) = -2,5812 + 0,0096x_{t1},$$

tal que x_{t1}, referente a variável duração da lactação, aumenta em $0,96\%$ a produção de leite.

3.2 As Distribuições Beta Generalizadas

Um dos maiores benefícios da classe de distribuições beta generalizadas é sua habilidade de ajustar dados assimétricos que possam não ser propriamente ajustados pelas distribuições usuais. Considere-se, inicialmente, a função de distribuição acumulada (f.d.a.) $G(x)$. Eugene et al. (2002) definiram uma classe de distribuições generalizadas a partir de $G(x)$, sendo esta expressa como

$$F(x) = \frac{1}{B(a,b)} \int_0^{G(x)} w^{a-1}(1-w)^{b-1} dw, \tag{3.3}$$

em que $a > 0$ e $b > 0$ são dois parâmetros adicionais cujo objetivo é introduzir assimetria e variar o peso das caudas e $B(a,b) = \int_0^1 w^{a-1}(1-w)^{b-1}dw$ é a função beta. A f.d.a. $G(x)$ poderia ser uma função arbitrária e F é nominada como distribuição beta G. A aplicação de $X = G^{-1}(V)$ para $V \sim beta(a,b)$ produz X seguindo (3.3).

Eugene et al. (2002) definiram a distribuição beta normal (BN) tomando a função $G(x)$ como a f.d.a. da distribuição normal e obteve alguns de seus primeiros momentos. Nadarajah e Kotz (2004) introduziram a distribuição beta Gumbel (BG) tendo como $G(x)$ a f.d.a. da distribuição Gumbel, provando que esta apresenta expressões em forma fechada para o cálculo dos momentos, a distribuição assintótica da estatística de ordem extrema e discutiram o processo de estimação dos parâmetros por máxima verossimilhança. Nadarajah e Gupta (2004) apresentaram a distribuição beta Fréchet (BF) a partir de $G(x)$ sendo a distribuição de Fréchet, obtiveram a forma analítica f.d.p. e a função da razão de risco e calcularam a distribuição assintótica da estatística de ordem extrema. Além disso, Nadarajah e Kotz (2005) trabalharam com a distribuição beta exponencial (BE) e obtiveram a função geradora de momentos, os primeiros quatro cumulantes, a distribuição assintótica da estatística de ordem extrema e discursaram sobre a estimação de máxima verossimilhança.

Pode-se reescrever (3.3) como

$$F(x) = I_{G(x)}(a,b), \tag{3.4}$$

em que $I_y(a,b) = B(a,b)^{-1} \int_0^y w^{a-1}(1-w)^{b-1}dw$ denota a razão da função beta incompleta, isto é, a f.d.a. da distribuição beta com parâmetros a e b. Para a e b gerais, pode-se expressar (3.4) em termos da função hipergeométrica conhecida que é definida por

$${}_2F_1(\alpha,\beta,\gamma;x) = \sum_{i=0}^{\infty} \frac{(\alpha)_i(\beta)_i}{(\gamma)_i i!} x^i,$$

em que $(\alpha)_i = \alpha(\alpha+1)\cdots(\alpha+i-1)$ denota o fatorial ascendente. Obtém-se

$$F(x) = \frac{G(x)^a}{aB(a,b)}\ {}_2F_1(a, 1-b, a+1; G(x)).$$

As propriedades de $F(x)$ ou qualquer outra distribuição beta G definidas a partir de uma família $G(x)$ em (3.3), poderia, em princípio, seguir das propriedades da função hipergeométrica as quais são bem estabelecidas na literatura; veja, por exemplo, a Seção 9.1 de Gradshteyn e Ryzhik (2000).

A f.d.p. correspondente a (3.3) pode ser escrita como

$$f(x) = \frac{g(x)}{B(a,b)} G(x)^{a-1}\{1-G(x)\}^{b-1}, \tag{3.5}$$

em que $g(x) = dG(x)/dx$ é a f.d.p. da distribuição original. A f.d.p. $f(x)$ será melhor empregada quando a f.d.a. e a f.d.p. $g(x) = dG(x)/dx$ apresentarem uma expressão analítica simples. Exceto para alguns casos especiais, a escolha de $G(x)$ em (3.3) implicaria que a f.d.p. seria difícil de ser aplicada.

Cordeiro e Cristino (2009) introduziram o quarto parâmetro à distribuição beta Rayleigh generalizada (BRG) tomando $G(x)$ em (3.3) como sendo a f.d.a. da distribuição Rayleigh generalizada (RG). Portanto, a f.d.a. da distribuição BRG é escrita como

$$F(x) = I_{\frac{\gamma(\alpha+1,\theta x^2)}{\Gamma(\alpha+1)}}(a,b), \quad a>0,\ b>0,\ \theta>0\ e\ \alpha>0, \tag{3.6}$$

e a f.d.p. associada é

$$f(x) = \frac{2\theta^{\alpha+1}x^{2\alpha+1}e^{-\theta x^2}}{B(a,b)\Gamma(\alpha+1)}\left\{\frac{\gamma(\alpha+1,\theta x^2)}{\Gamma(\alpha+1)}\right\}^{a-1}\left\{1-\frac{\gamma(\alpha+1,\theta x^2)}{\Gamma(\alpha+1)}\right\}^{b-1}, \tag{3.7}$$

em que $\gamma(\alpha,x) = \int_0^x w^{\alpha-1}e^{-w}dw$ é a função gama incompleta.

Se X é uma variável com f.d.p. (3.7), escreve-se $X \sim BRG(a,b,\lambda,\alpha)$. A distribuição BRG estende algumas distribuições conhecidas. A distribuição Rayleigh generalizada é um caso especial quando $a=b=1$. Além disso, quando $\alpha=0$, a distribuição Rayleigh é obtida. A distribuição Rayleigh generalizada exponencializada (RGE) corresponde a $b=1$.

Barreto-Souza et al. (2009) apresentaram a distribuição beta exponencial generalizada (BEG) com quatro parâmetros em que $G(x)$ em (3.3) corresponde a f.d.a. da distribuição exponencial generalizada. A f.d.a. da distribuição BEG é, então,

$$F(x) = \frac{1}{B(a,b)}\int_0^{(1-e^{-\lambda x})^\alpha} x^{a-1}(1-x)^{b-1}dx, \quad x>0, \tag{3.8}$$

para $a > 0, b > 0, \lambda > 0$ e $\alpha > 0$. A f.d.p. da nova distribuição é especificada como

$$f(x) = \frac{\alpha\lambda}{B(a,b)}\, e^{-\lambda x}(1-e^{-\lambda x})^{\alpha a-1}\{1-(1-e^{-\lambda x})^{\alpha}\}^{b-1}, \quad x > 0. \tag{3.9}$$

A função densidade (3.9) não envolve qualquer função complicada. Se X é uma variável aleatória com f.d.p. (3.9), escreve-se $X \sim BEG(a,b,\lambda,\alpha)$. A distribuição BEG generaliza algumas distribuições convencionais da literatura. A distribuição generalizada exponencial é um caso especial quando escolhido $a = b = 1$. Se adicionar $\alpha = 1$, obtém-se a distribuição exponencial com parâmetro λ. A distribuição beta exponencial é obtida de (3.8) com $\alpha = 1$. A distribuição generalizada exponencial dupla (GED) corresponde a $a = 1$ em (3.8).

É muito importante conhecer a aplicabilidade das distribuições de probabilidade referentes às distribuições beta generalizadas. Ressalta-se, contudo, que, devido a existência de estudos referentes à densidade da maioria das distribuições aqui abordadas, apresenta-se, aqui, apenas a densidade correspondente de cada distribuição beta generalizada.

A princípio, tem-se que uma variável aleatória X segue uma distribuição beta normal $BN(a,b,\mu,\sigma^2)$ se sua f.d.p. é expressa como

$$f(x) = \frac{\sigma^{-1}}{B(a,b)}\phi\left(\frac{x-\mu}{\sigma}\right)\left\{\Phi\left(\frac{x-\mu}{\sigma}\right)\right\}^{a-1}\left\{1-\Phi\left(\frac{x-\mu}{\sigma}\right)\right\}^{b-1},$$

em que $x \in \mathbb{R}$, $\mu \in \mathbb{R}$ é um parâmetro de locação, $\sigma > 0$ é um parâmetro de escala, a e b são parâmetros de forma, e $\phi(.)$ e $\Phi(.)$ são a f.d.p. e f.d.a. da distribuição normal padrão, respectivamente. Para $\mu = 0$ e $\sigma = 1$, obtém-se a distribuição beta normal (padrão). Por muitas décadas, a distribuição normal teve uma posição central na estatística e, mediante isto, tornou-se a base de muitos trabalhos práticos estatísticos, particularmente na astronomia. Segundo Eugene et al. (2002), a distribuição beta normal apresenta grande flexibilidade não apenas na modelagem de distribuições simétricas com caudas pesadas, mas, também, na modelagem de distribuições assimétricas e bimodais. Mediante isto, acredita-se que, possivelmente, o futuro dos modelos de regressão na estatística passará pela distribuição beta normal devido a sua maior flexibilidade.

No caso em que a variável aleatória X segue uma distribuição beta gama $BG(a,b,\alpha,\beta)$, tem-se que sua f.d.p. é dada por

$$f(x) = \frac{\beta^{\alpha}x^{\alpha-1}e^{-\beta x}}{B(a,b)\Gamma(\alpha)^{a+b-1}}\gamma(\alpha,\beta x)^{a-1}\{\Gamma(\alpha)-\gamma(\alpha,\beta x)\}^{b-1}, \quad x > 0,$$

em que $G(x) = \gamma(\alpha,\beta x)/\Gamma(\alpha)$ é a f.d.a. da distribuição gama com parâmetros α e β. Pode ser visto em Kong et al. (2007) que a distribuição gama e suas distribuições generalizadas

têm sido muito aplicadas à análise de distribuição de renda, testes de sobrevivência, e muitos fenômenos físicos e econômicos. Segundo Johnson et al. (1995a, cap. 17) existem ainda outros tópicos em que é possível utilizar a distribuição gama associada com o processo aleatório no tempo, em particular, no processo de precipitação meteorológica. Tem-se, ainda, que Dennis e Patil (1984) citam aplicações da distribuição gama em estudos estatísticos baseados em ecologia.

Seja X uma variável aleatória que segue a distribuição beta Gumbel, a sua f.d.p. é expressada por

$$f(x) = \frac{ue^{-au}\{1-e^{-u}\}^{b-1}}{\beta B(a,b)},$$

$-\infty < x, \alpha < \infty$ e $a, b, \beta > 0$, em que $u = \exp\{-(x-\alpha)/\beta\}$ (NADARAJAH; KOTZ, 2004). A distribuição Gumbel e suas generalizações são talvez as distribuições estatísticas mais aplicadas para problemas na engenharia. Algumas de suas recentes áreas de aplicação incluem análise de freqüência de inundações, engenharia de sistemas, engenharia nuclear, engenharia de pesca, engenharia de risco-base, engenharia espacial, engenharia de confiabilidade de software, engenharia estrutural e engenharia dos ventos. No livro de Kotz e Nadarajah (2000) existem listas correspondentes a cinquenta aplicações desde aceleração de vida até testes de terremotos, inundações, corridas de cavalos, precipitações, "queues" em mercados, correntes marinhas, velocidade do vento e registros de circuitos de corridas (apenas para mencionar algumas).

Seja X uma variável aleatória que segue a distribuição beta log-normal $BLN(a, b, \mu, \sigma^2)$, tem-se sua f.d.p. expressa por

$$f(x) = \frac{e^{-\frac{\{\ln(x)-\mu\}^2}{2\sigma^2}}}{xB(a,b)(\sigma\sqrt{2\pi})^{a+b-1}} \left\{ \int_0^x \frac{e^{-\frac{\{\ln(t)-\mu\}^2}{2\sigma^2}}}{t} dt \right\}^{a-1} \left\{ \sigma\sqrt{2\pi} - \int_0^x \frac{e^{-\frac{\{\ln(t)-\mu\}^2}{2\sigma^2}}}{t} dt \right\}^{b-1}.$$

Observe que a distribuição log-normal é utilizada para valores de x positivos, $a, b > 0$, $\mu \in \mathbb{R}$ e $\sigma > 0$. A distribuição log-normal pode ser aplicada a dados econômicos, particularmente, a função de produção, sendo esta denominada, às vezes, de distribuição *Cobb-Douglas*. A distribuição log-normal e, por conseguinte, a distribuição beta log-normal podem ser aplicadas, ainda, na agricultura, entomologia e, também, numa variedade de situações biológicas e farmacológicas (JOHNSON et al., 1995a, cap. 14). Além de, também, poderem ser aplicadas a dados de sobrevivência.

A distribuição F resulta frequentemente como distribuição nula de uma estatística de teste, especialmente nos testes da razão de verosimilhanças, talvez sendo a mais notável nas análises de variância. Considere a distribuição $F(2\alpha, 2\beta)$ com g.l. 2α e 2β para $x > 0$,

$\alpha > 0$, and $\beta > 0$, e f.d.p. e f.d.a. expressa como

$$g(x) = \frac{\alpha^{\alpha} x^{\alpha-1}}{\beta^{\alpha} B(\alpha,\beta)(1+\alpha x/\beta)^{\alpha+\beta}}$$

e $G(x) = I_{\frac{\alpha x}{\alpha x+\beta}}(\alpha,\beta)$, respectivamente. Seja, ainda, X uma variável aleatória que segue a distribuição beta F, a correspondente f.d.p. da distribuição beta F $BF(a,b,2\alpha,2\beta)$ com parâmetros $a,b,2\alpha$ and 2β, para qualquer $x > 0$, pode ser escrita como

$$f(x) = \frac{\alpha^{\alpha} x^{\alpha-1}}{\beta^{\alpha} B(\alpha,\beta)(1+\alpha x/\beta)^{\alpha+\beta} B(a,b)} I_{\frac{\alpha x}{\alpha x+\beta}}(\alpha,\beta)^{a-1} \left\{1 - I_{\frac{\alpha x}{\alpha x+\beta}}(\alpha,\beta)\right\}^{b-1}.$$

A distribuição t-Student é a segunda distribuição mais popular na estatística, perdendo apenas para a distribuição normal (CORDEIRO; NADARAJAH, 2009). A f.d.p. da distribuição t-Student t_ν com g.l. $\nu > 0$ (para $-\infty < x < \infty$) é

$$g(x) = \frac{1}{\sqrt{\nu} B(1/2,\nu/2)} \left(1 + \frac{x^2}{\nu}\right)^{-(\nu+1)/2}.$$

Para qualquer x real, a f.d.a. da distribuição t-Student t_ν é expressa como $G(x) = I_y(1/2, \nu/2)$, em que $y = (x+\sqrt{x^2+\nu})/(2\sqrt{x^2+\nu})$. A f.d.p. da distribuição beta Student $BS(a,b,\nu)$ com parâmetros ν, a e b é, portanto, é escrita como (para qualquer x)

$$f(x) = \frac{\left(1+\frac{x^2}{\nu}\right)^{-(\nu+1)/2}}{\sqrt{\nu} B(a,b) B(1/2,\nu/2)} \left\{I_{\frac{x+\sqrt{x^2+\nu}}{2\sqrt{x^2+\nu}}}(1/2,\nu/2)\right\}^{a-1} \left\{1 - I_{\frac{x+\sqrt{x^2+\nu}}{2\sqrt{x^2+\nu}}}(1/2,\nu/2)\right\}^{b-1}. \tag{3.10}$$

A distribuição beta Student será simétrica em torno de zero apenas quando $a = b$.

A f.d.p. e f.d.a. da distribuição beta $B(\alpha,\beta)$ com parâmetros $\alpha > 0$ e $\beta > 0$ são, respectivamente, $g(x) = x^{\alpha-1}(1-x)^{\beta-1}/B(\alpha,\beta)$ e $G(x) = I_x(\alpha,\beta) = B(\alpha,\beta)^{-1} \int_0^x t^{\alpha-1}(1-t)^{\beta-1} dt$ para $0 < x < 1$. Seja, ainda, X uma variável aleatória que segue a distribuição beta beta $BB(a,b,\alpha,\beta)$ é expressa como

$$f(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)B(a,b)} I_x(\alpha,\beta)^{a-1} \{1 - I_x(\alpha,\beta)\}^{b-1}. \tag{3.11}$$

No que corresponde às distribuições beta F, beta Student e beta beta, existem estudos recentes em que se pretende compreender melhor a funcionalidade dessas distribuições. Pode-se, porém, afirmar que estas são abordadas quando tratar-se de dados com caudas pesadas. O estudo gráfico da f.d.p. das diferentes distribuições comparando suas formas segue processo análogo mostrado no Capítulo 4 para a distribuição beta power.

3.2.1 Estudo Numérico das Distribuições Beta Generalizadas

Utiliza-se o estudo numérico de diferentes distribuições a fim de analizar a potencialidade das distribuições beta generalizadas com respeito às distribuições usuais. As Tabelas 3.5 a 3.10 a seguir apresentam o cálculo dos momentos ordinários para as distribuições beta normal, $BN(a,b,0,1)$, beta gama, $BG(a,b,2,3)$, beta beta, $BB(a,b,2,3)$, beta Student, $BS(a,b,6)$, beta F, $BF(a,b,4,6)$, beta Gumbel, $BGB(a,b,2,4)$ e beta lognormal, $BLN(a,b,0,1)$. Os casos particulares ocorrem quando $a=b=1$ para cada distribuição abordada, ou seja, seriam obtidas as distribuições normal, gama, beta, t-Student, F, Gumbel e log-normal. Nota-se, nas Tabelas de 3.5 a 3.10, que ao ser utilizar as distribuições beta generalizadas tanto a assimetria quanto a curtose apresentam melhoras com relação aos casos especiais de cada distribuição aqui apresentada para as correspondentes parametrizações abordadas.

As Figuras 3.2 a 3.15 ilustram o comportamento de assimetria e curtose das distribuições beta generalizadas aqui abordadas para o caso em que a é um parâmetro fixo com b sendo um parâmetro escolhido o intervalo $[1,6]$, e também para o caso em que b é um parâmetro fixo com a sendo um parâmetro escolhido para o mesmo intervalo.

As medidas de variância, assimetria (α_3) e curtose (α_4) para as distribuições abordadas são calculadas usando as relações

$$\sigma^2 = Var(X) = E(X^2) - E(X)^2,$$
$$\alpha_3 = \frac{E(X^3) - 3E(X)E(X^2) + 2E(X)^3}{Var(X)^{3/2}} \quad \text{e}$$
$$\alpha_4 = \frac{E(X^4) - 4E(X)E(X^3) + 6E(X^2)E(X)^2 - 3E(X)^4}{Var(X)^2}.$$

Essas relações podem ser verificadas em Nadarajah e Kotz (2004). E o estudo numérico tanto para os momentos até sexta ordem quanto para a assimetria e curtose foi realizado através dos softwares MATLAB na versão 7.3 (R2006b) e MAPLE na versão 11.0, juntamente com o pacote R na versão 2.8.0 para a formação gráfica.

O algoritmo utilizado para tais cálculos é apresentado para a distribuição beta normal, sendo o método análogo para as demais distribuições beta generalizadas conforme as parametrizações de cada distribuição e sua estrutura encontra-se no Apêndice A.

Ao analisar os gráficos de assimetria da $BN(a,b,0,1)$, correspondentes às Figuras 3.2 e 3.3, tem-se que para a fixo, a assimetria decresce enquanto, para b fixo, esta cresce. Em se tratando da curtose, para a fixo, ocorre um decréscimo para valores de b menores do

que dois. E, para valores de b a partir de dois, a curtose passa a crescer. O mesmo ocorre para o caso em que $BN(a,b,0,1)$ é função de b para valores fixos de a.

Para as Figuras 3.4 e 3.5, referentes a $BG(a,b,2,3)$, observa-se que, quando fixado a, a assimetria e curtose decrescem com exceção do caso em que $a = 3,5$. E, quando fixado b, a assimetria e curtose decrescem.

No caso da distribuição $BB(a,b,2,3)$, através das Figuras 3.6 e 3.7, observa-se que a assimetria é crescente quando se utiliza como função de b e fixado a. E, ainda, esta decresce quando se faz função de a e fixa b. Para o caso da curtose, quando se fixa a, ocorre um crescimento da mesma. E, em se tratando de b fixo, existe um decréscimo até valores de b menores que três, sendo que para $b = 1,0$ e $1,5$, a curtose já passa a crescer.

Para a distribuição $BS(a,b,6)$, através das Figuras 3.8 e 3.9, observa-se que a assimetria decresce como função de b e a fixado. E, cresce para função de a e b fixado. No caso da curtose, quando se tem como função de b, ocorre um decréscimo, exceto para $a = 1,0$, pois este passa a crescer quando $b > 1,25$ de forma notória. E, para $a = 1,5$, também apresenta-se um crescimento quando b assume valores maiores que dois. Essa mesma situação é observada quando fixado b, para valores de a no intervalo $[1,6]$.

As Figuras 3.10 e 3.11, correspondentes a $BF(a,b,4,6)$, mostram que, para o caso de assimetria, há um decréscimo seja para a ou b fixos. E, para a curtose, quando a fixo, ocorre um breve salto quando $a = 2,5$ e $b \leq 1,25$. Para os demais casos de a fixo, apresenta-se um decrescimento da curtose. Para o caso de b fixo, na Figura 3.11, o decréscimo ocorre para $b = 1,0$ e, para os demais valores, a curtose permanece aparentemente constante.

A distribuição $BLN(a,b,0,1)$ apresenta decréscimo tanto na assimetria quanto na curtose para ambos os casos de a e b fixos. Sendo interessante notar, através das Figuras 3.12 e 3.13, que para o caso de a fixo, o decréscimo da assimetria e da curtose é mais acelerado.

Para a distribuição $BGB(a,b,2,4)$, observa-se, na Figura 3.14, que a assimetria cresce enquanto a curtose decresce. Na Figura 3.15, a assimetria decresce lentamente apresentando uma aparente constância para valores de $b = 2,5$ e $3,5$. E, no caso da curtose, apresenta-se um crescimento lento quanto $b = 2,5$ e $3,5$ com uma aparente constância para os demais valores assumidos por b. Nota-se, ainda, que para $b = 1,0$ a assimetria e a curtose permanecem constantes.

Tabela 3.5: Momentos ordinários para diferentes distribuições considerando $a = 1,0$ e $b = 1,0$.

μ'_k	$BN(0,1)$	$BG(2,3)$	$BB(2,3)$	$BS(6)$	$BF(4,6)$	$BGB(2,4)$	$BLN(0,1)$
μ'_1	0	6,00000	0,40000	0	1,50000	-0,30886	1,64870
μ'_2	1	48,00000	0,20000	1,5	6,70960	26,41430	7,38900
μ'_3	0	480,00000	0,11429	0	179,27080	-178,27960	90,01290
μ'_4	3	5760,0000	0,07143	13,49980	3,89 $\times 10^4$	3945,67300	2975,53220
μ'_5	0	80640,000	0,04762	0	1,99 $\times 10^7$	-71909,78470	2,61$\times 10^5$
μ'_6	15	1,29$\times 10^6$	0,03333	2094,53850	1,33$\times 10^{10}$	1.79$\times 10^6$	5,37$\times 10^7$
σ^2	1	12,00000	0,04000	1,5	4,45950	26,31890	4,67100
α_3	0	1,15470	0,28571	0	16,54660	-1,13950	6,18410
α_4	3	5,00000	2,35710	5,99999	1906,38730	5,40000	113,68060

Tabela 3.6: Momentos ordinários para diferentes distribuições considerando $a = 1,0$ e $b = 1,5$.

μ'_k	$BN(0,1)$	$BG(2,3)$	$BB(2,3)$	$BS(6)$	$BF(4,6)$	$BGB(2,4)$	$BLN(0,1)$
μ'_1	-0,34405	4,79760	0,32971	-0,40954	0,95622	1,46440	1,04490
μ'_2	0,91189	29,90150	0,13828	1,33860	1,88420	17,50070	2,30670
μ'_3	-0,91686	228,36750	0,06736	-2,29620	8,13670	17,41850	10,43860
μ'_4	2,56460	2060,55590	0,03640	10,88910	116,53120	958,57490	95,11410
μ'_5	-4,12670	21421.58550	0,02124	-63,41610	1,18$\times 10^4$	-4556,49590	1726,54400
μ'_6	12,26950	2,52$\times 10^5$	0,01315	1578,47050	4,58$\times 10^6$	1,40$\times 10^5$	62021,66720
σ^2	0,79352	6,88450	0,02956	1,17090	0,96982	15,35620	1,21490
α_3	-0,08079	1,04370	0,44574	-0,62272	4,69110	-0,88381	4,09910
α_4	3,03080	4,60370	2,63260	6,12010	99,1315	4,52870	42,69490

Tabela 3.7: Momentos ordinários para diferentes distribuições considerando $a = 1,0$ e $b = 3,5$.

μ'_k	$BN(0,1)$	$BG(2,3)$	$BB(2,3)$	$BS(6)$	$BF(4,6)$	$BGB(2,4)$	$BLN(0,1)$
μ'_1	-0,94600	3,12430	0,21430	-1,11950	0,45985	3,99060	0,49691
μ'_2	1,41640	12,15100	0,05953	2,21200	0,33507	22,13010	0,38758
μ'_3	-2,41740	55,96840	0,01961	-5,87610	0,35475	130,96860	0,45208
μ'_4	4,79860	296,15280	0,00731	22,76940	0,52534	869,82350	0,76425
μ'_5	-10,50730	1763,52640	0,00300	-151,90660	1,07690	6018,35720	1,83450
μ'_6	25,21950	11643,03290	0,00133	3661,49290	3,08620	44831,57450	6,16320
σ^2	0,52150	2,38970	0,01360	0,95881	0,12360	6,20520	0,14066
α_3	-0,24099	0,83181	0,64546	-1,33480	2,00180	-0,44438	2,26880
α_4	3,14050	3,94190	3,13700	9,11340	10,72050	3,45310	12,98770

Tabela 3.8: Momentos ordinários para diferentes distribuições considerando $a = 1,5$ e $b = 1,5$.

μ'_k	$BN(0,1)$	$BG(2,3)$	$BB(2,3)$	$BS(6)$	$BF(4,6)$	$BGB(2,4)$	$BLN(0,1)$
μ'_1	0	5,75540	0,39522	0	1,24590	0,015559	1,36530
μ'_2	0,62203	40,23850	0,18340	0,81558	2,84110	15,16670	3,49440
μ'_3	0	331,38480	0,09517	0	13,26520	-52,37660	16,98780
μ'_4	1,17200	3145,05090	0,05365	2,79680	196,48700	1047,42530	159,39160
μ'_5	0	33834,33820	0,03225	0	$2,00\times10^4$	-11526,94250	2921,69100
μ'_6	3,71060	$4,07\times10^5$	0,02040	26,69000	$7,78\times10^6$	$2,07\times10^5$	$1,05\times10^5$
σ^2	0,62203	7,11410	0,02720	0,81558	1,28890	15,16640	1,63020
α_3	0	0,94381	0,26416	0	4,45180	-0,89876	3,73060
α_4	3,02920	4,37940	2,54910	4,20460	90,06360	4,56790	35,84980

Tabela 3.9: Momentos ordinários para diferentes distribuições considerando $a = 1,5$ e $b = 2,5$.

μ'_k	$BN(0,1)$	$BG(2,3)$	$BB(2,3)$	$BS(6)$	$BF(4,6)$	$BGB(2,4)$	$BLN(0,1)$
μ'_1	-0,38915	4,47390	0,31334	-0,43863	0,79106	1,88760	0,85339
μ'_2	0,62203	23,75800	0,11655	0,81558	0,95754	11,84310	1,13980
μ'_3	-0,63769	145,65200	0,04903	-1,12010	1,72890	38,48180	2,34580
μ'_4	1,17200	1010,94950	0,02266	2,79680	4,72200	338,28130	7,38100
μ'_5	-1,78510	7829,76040	0,01128	-7,15760	20,72110	948,02150	35,3721
μ'_6	3,71060	66909,75110	0,00597	26,69000	170,66540	16077,76050	257,78310
σ^2	0,47059	3,74240	0,01837	0,62319	0,33177	8,28000	0,41152
α_3	-0,09098	0,81158	0,40319	-0,43840	2,33660	-0,63515	2,54060
α_4	3,05170	3,98000	2,79380	4,27950	15,18920	3,83370	16,3143

Tabela 3.10: Momentos ordinários para diferentes distribuições considerando $a = 2,5$ e $b = 3,5$.

μ'_k	$BN(0,1)$	$BG(2,3)$	$BB(2,3)$	$BS(6)$	$BF(4,6)$	$BGB(2,4)$	$BLN(0,1)$
μ'_1	-0,24014	4,79300	0,33919	-0,26180	0,84743	1,38030	0,90936
μ'_2	0,35014	25,57150	0,12786	0,42072	0,94008	7,45550	1,10100
μ'_3	-0,23245	150,18740	0,05248	-0,33435	1,34960	18,69580	1,76890
μ'_4	0,37198	962,54130	0,02311	0,62776	2,50610	139,26640	3,76700
μ'_5	-0,38132	6683,61800	0,01080	-0,85345	6,07910	327,42030	10,6424
μ'_6	0,66552	49984,27710	0,00531	1,90960	19,7062	4238,71880	39,9764
σ^2	0,29247	2,59890	0,01281	0,35218	0,22194	5,55030	0,27405
α_3	-0,04996	0,64709	0,28978	-0,19044	1,69090	-0,52895	1,87680
α_4	3,03800	3,64320	2,82400	3,51970	8,82650	3,58300	9,90640

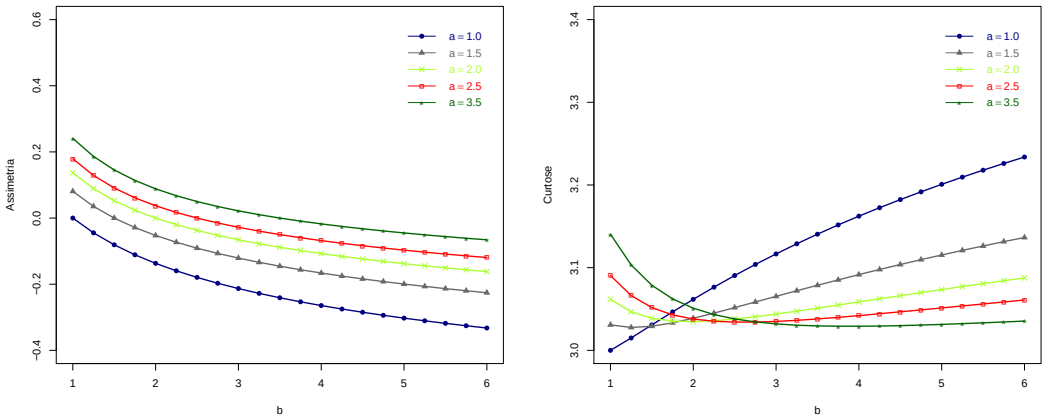

Figura 3.2: Gráficos de assimetria e curtose para a distribuição $BN(a,b,0,1)$ como função de b e fixado a.

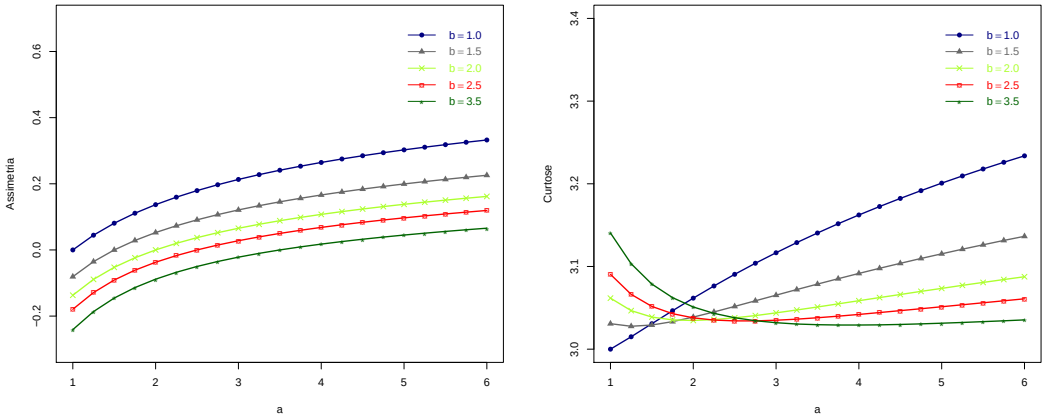

Figura 3.3: Gráficos de assimetria e curtose para a distribuição $BN(a,b,0,1)$ como função de a e fixado b.

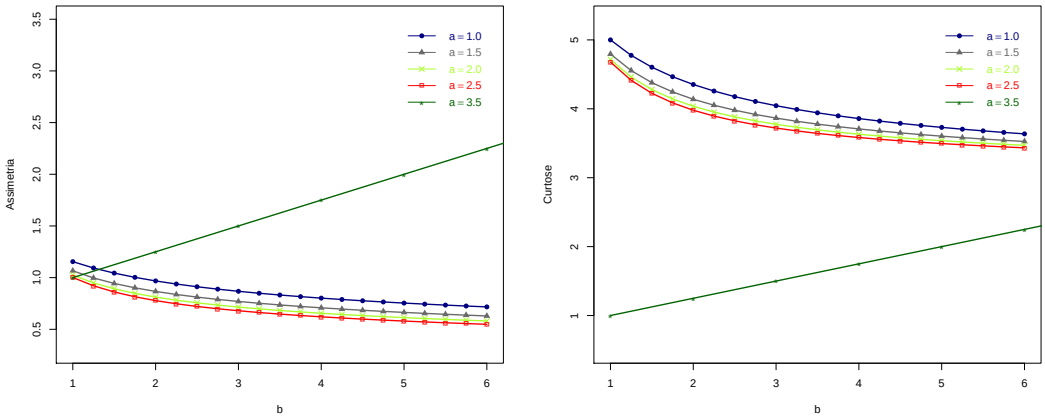

Figura 3.4: Gráficos de assimetria e curtose para a distribuição $BG(a,b,2,3)$ como função de b e fixado a.

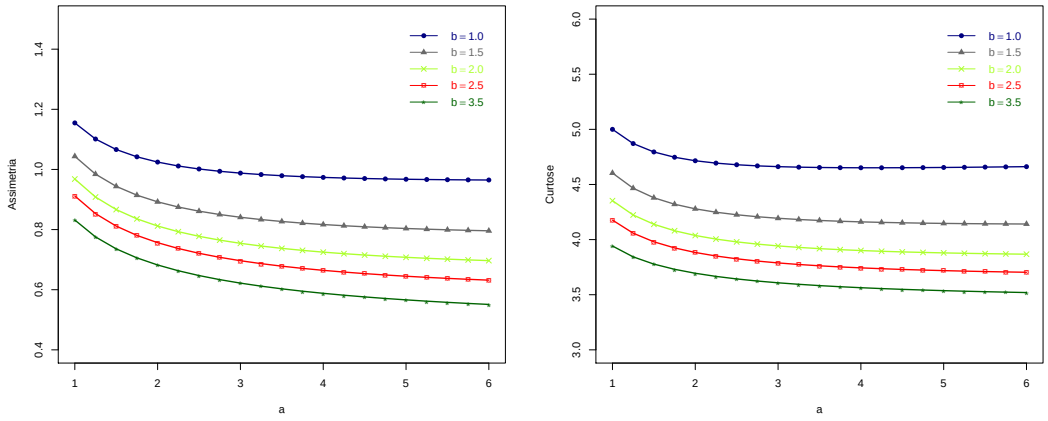

Figura 3.5: Gráficos de assimetria e curtose para a distribuição $BG(a,b,2,3)$ como função de a e fixado b.

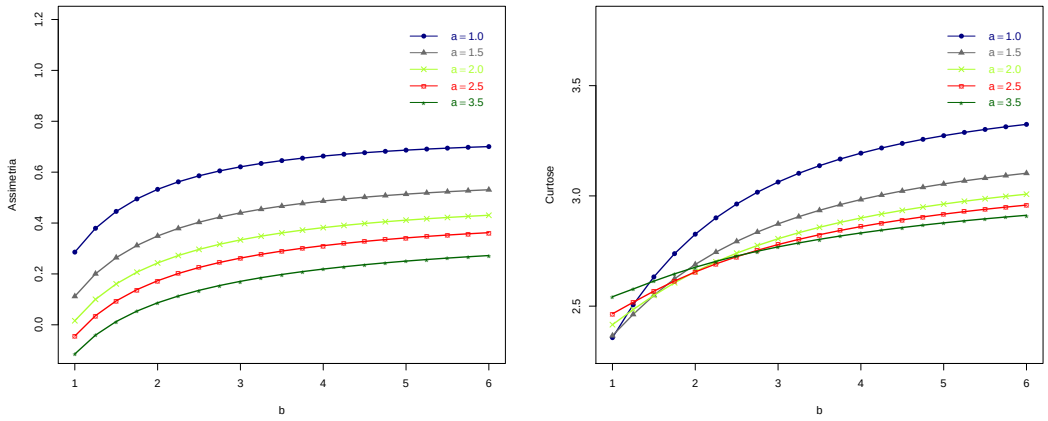

Figura 3.6: Gráficos de assimetria e curtose para a distribuição $BB(a,b,2,3)$ como função de b e fixado a.

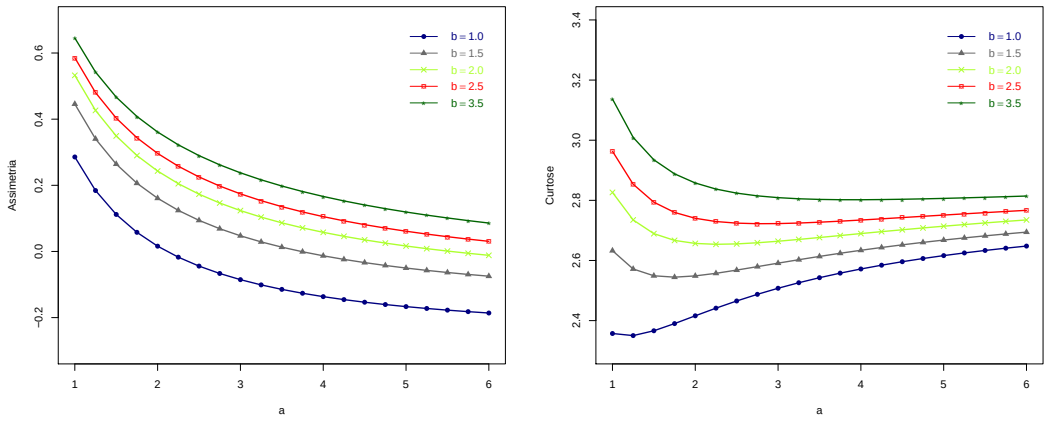

Figura 3.7: Gráficos de assimetria e curtose para a distribuição $BB(a,b,2,3)$ como função de a e fixado b.

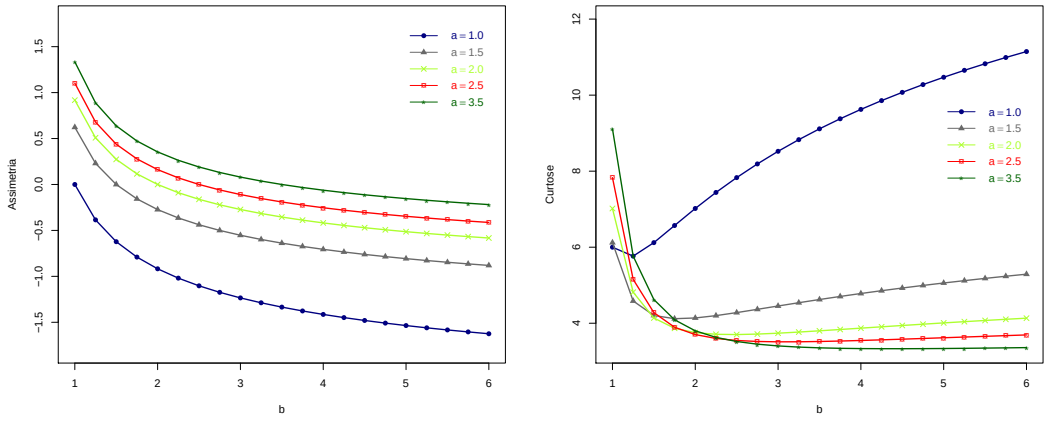

Figura 3.8: Gráficos de assimetria e curtose para a distribuição $BS(a,b,6)$ como função de b e fixado a.

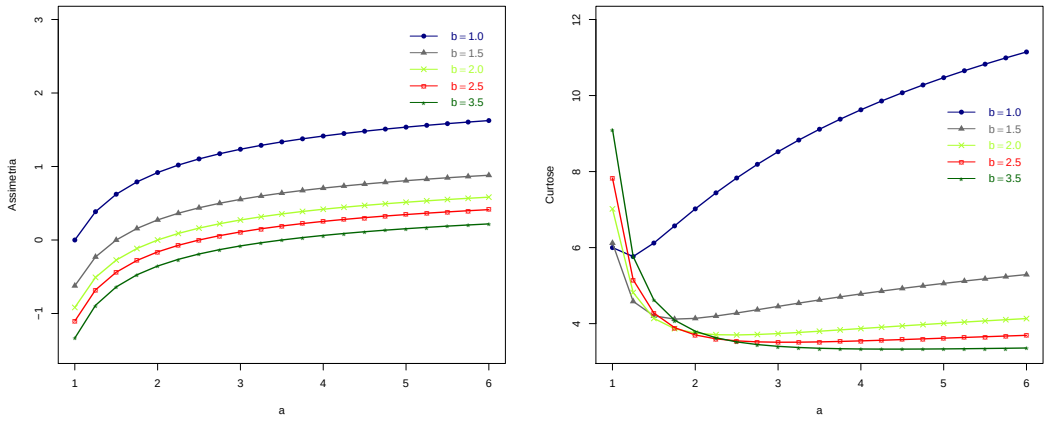

Figura 3.9: Gráficos de assimetria e curtose para a distribuição $BS(a,b,6)$ como função de a e fixado b.

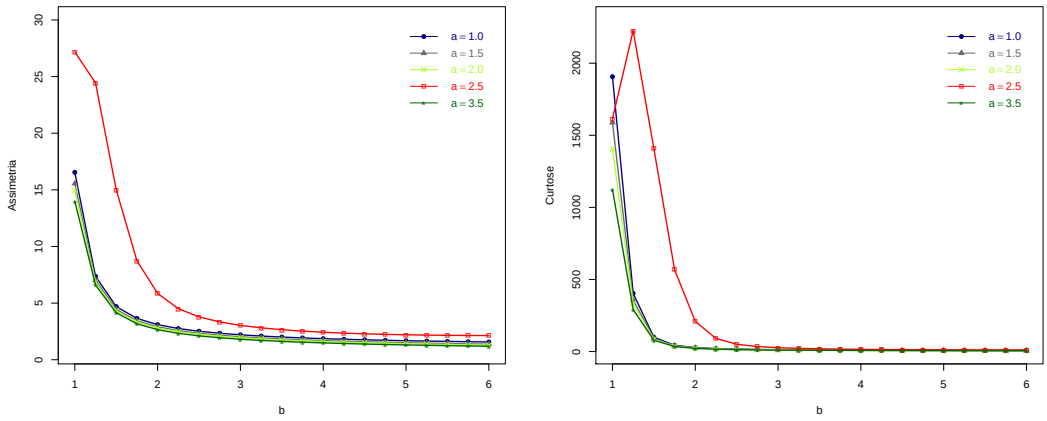

Figura 3.10: Gráficos de assimetria e curtose para a distribuição $BF(a,b,4,6)$ como função de b e fixado a.

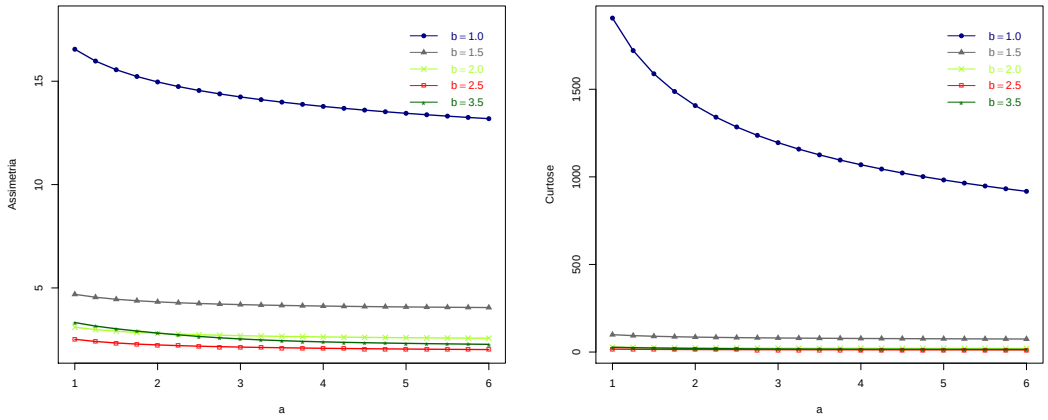

Figura 3.11: Gráficos de assimetria e curtose para a distribuição $BF(a,b,4,6)$ como função de a e fixado b.

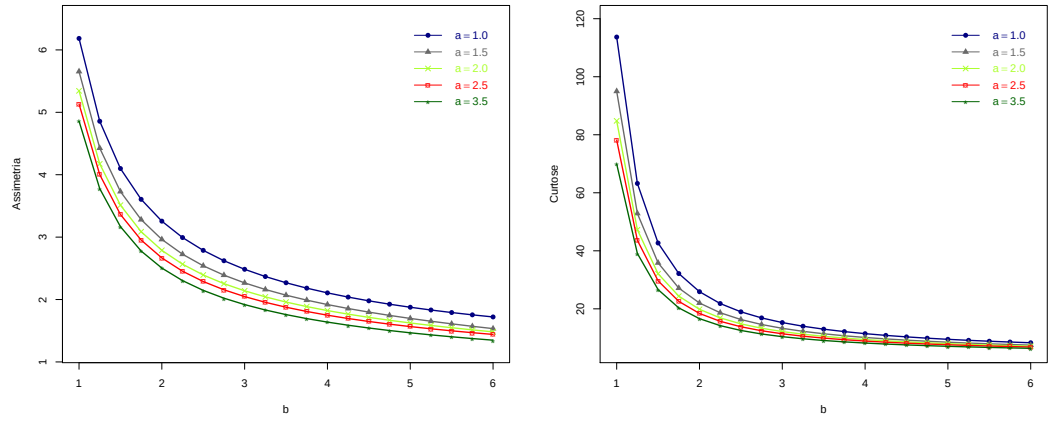

Figura 3.12: Gráficos de assimetria e curtose para a distribuição $BLN(a,b,0,1)$ como função de b e fixado a.

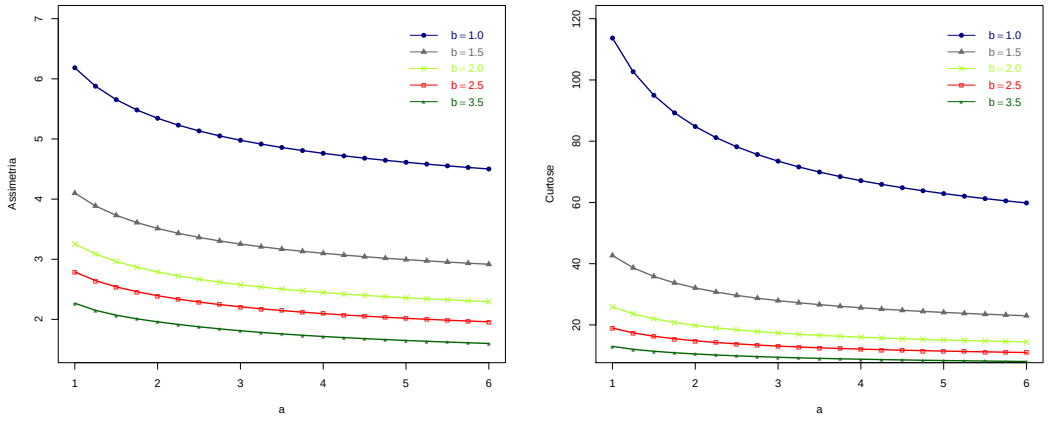

Figura 3.13: Gráficos de assimetria e curtose para a distribuição $BLN(a,b,0,1)$ como função de a e fixado b.

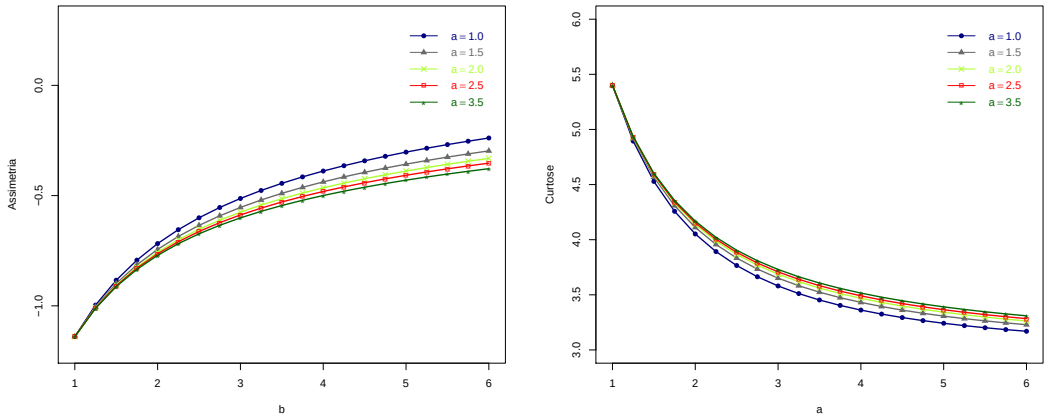

Figura 3.14: Gráficos de assimetria e curtose para a distribuição $BGB(a,b,2,4)$ como função de b e fixado a.

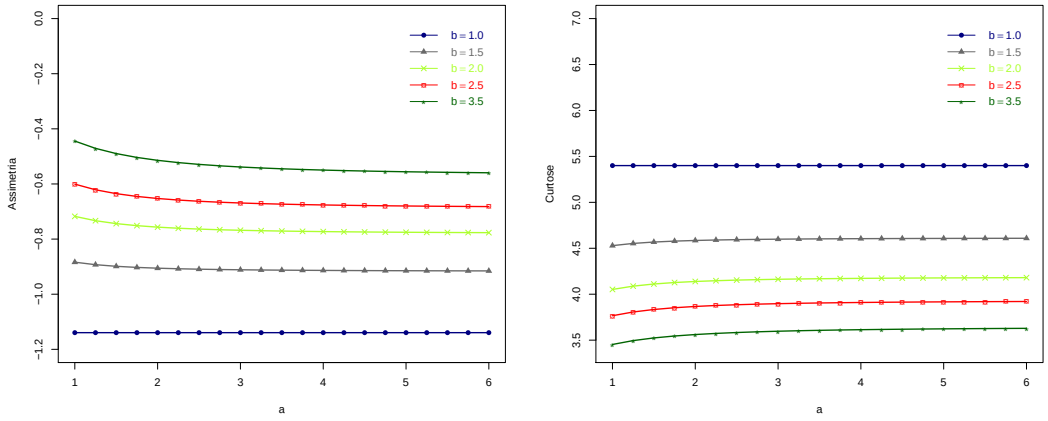

Figura 3.15: Gráficos de assimetria e curtose para a distribuição $BGB(a,b,2,4)$ como função de a e fixado b.

3.2.2 Estudo Numérico considerando a função de Lauricella

As funções de Lauricella são generalizações das funções hipergeométricas de Gauss em múltiplas variáveis. Essas generalizações foram investigadas por Lauricella (1893) e mais profundamente por Appel e Fériet (1926, p. 117). E as propriedades das funções hipergeométricas podem ser econtradas em Prudnikov et al. (1992), Gradshteyn e Ryzhik (2000) e Nadarajah e Kotz (2009). Existe uma extensa literatura matemática direcionada ao estudo dessas funções: Srivastava desenvolveu a expansão polinomial das funções de Lauricella usando expansões envolvendo as funções de Bessel (SRIVASTAVA, 1991). Fórmulas integrais utilizando a função de Lauricella foram estabelecidas por Sharma e Singh (1990). As funções de Lauricella têm, também, interpretações probabilísticas, as quais resultam da avaliação de certos produtos de momentos de algumas distribuições multivariadas. Mostra-se que esses produtos de momentos podem ser expressados como

transformações através das funções de Lauricella (ONG; LEE, 2000). Outros problemas probabilísticos e estatísticos foram estudados em Mathai e Saxena (1987). E maiores detalhes sobre demais funcionalidades das funções de Lauricella podem ser encontrados em López e Ferreira (2003).

Seja n o número de variáveis, computa-se a função de Lauricella do tipo A usando a fórmula contida em Erdélyi (1936, p. 696, equação (1)) cuja expressão é dada por

$$F_A^{(n)}\left[\alpha,\beta_1,\ldots,\beta_n;\gamma_1,\ldots,\gamma_n;x_1,\ldots,x_n\right] = \frac{1}{\Gamma(\alpha)}\int_0^\infty t^{\alpha-1}\exp(-t)\ {}_1F_1\left(\beta_1;\gamma_1;x_1t\right)\cdots\ {}_1F_1\left(\beta_n;\gamma_n;x_nt\right)dt.$$

Conforme dito anteriormente, na Seção 3.2, as propriedades da f.d.a. $F(x)$ poderia, em princípio, seguir das propriedades da função hipergeométrica. Dessa forma, sendo a função de Lauricella uma generalização desta, faz-se o estudo da função de Lauricella para o caso em que

$$F_A^{(p)}\Big(\frac{p+k+1}{2};\frac{1}{2},\ldots,\frac{1}{2};\frac{3}{2},\ldots,\frac{3}{2};-1,\ldots,-1\Big).$$

Tabela 3.11: Cálculo dos momentos de uma distribuição empírica utilizando a função Lauricella para $p=3,5,8$ e 10, $\alpha^{(p)}=\frac{p+k+1}{2}$, $\beta_i=0,5$, $\gamma_i=1,5$, $x_i=-1$.

k	$\alpha^{(3)}$	$F_A^{(3)}$	$\alpha^{(5)}$	$F_A^{(5)}$	$\alpha^{(8)}$	$F_A^{(8)}$	$\alpha^{(10)}$	$F_A^{(10)}$
1	2,5	0,24030	3,5	0,05683	5,0	0,00405	6,0	0,00055
3	3,5	0,15348	4,5	0,02989	6,0	0,00171	7,0	0,00020
5	4,5	0,10436	5,5	0,01697	7,0	0,00079	8,0	0,00008
7	5,5	0,07495	6,5	0,01033	8,0	0,00039	9,0	0,00004
9	6,5	0,05636	7,5	0,00669	9,0	0,00021	10,0	0,00002
11	7,5	0,04403	8,5	0,00457	10,0	0,00012	11,0	$9{,}48\times10^{-6}$
19	11,5	0,02121	12,5	0,00145	14,0	0,00002	15,0	$1{,}24\times10^{-6}$

A Tabela 3.11 apresenta os resultados numéricos da análise correspondente em que se propõe o uso dessa generalização para a abordagem das distribuições beta generalizadas. O algoritmo para cálculo dos momentos de uma distribuição empírica utilizando a função Lauricella para $p=3$, $\alpha^{(3)}=11,5$, $\beta_i=0,5$, $\gamma_i=1,5$, $x_i=-1$ e $k=19$ é apresentado na Figura 3.16, sendo utilizado para cálculo o software MATHEMATICA®. Para as demais parametrizações de p é preciso atenção na extensão da função FA, definida para cálculo dos momentos. De modo análogo, é possível obter os resultados apresentados na Tabela 3.12.

```
FA = Function[{alpha, beta1, beta2, beta3, gamma1, gamma2, gamma3, x1, x2, x3},
      (Gamma[alpha])^(-1)*
      NIntegrate[Exp[-t]*t^(alpha - 1)*Hypergeometric1F1[beta1, gamma1, x1*t]*
              Hypergeometric1F1[beta2, gamma2, x2*t]*
              Hypergeometric1F1[beta3, gamma3, x3*t], {t, 0, Infinity}]]

N[FA[11.5, 0.5, 0.5, 0.5, 1.5, 1.5, 1.5, -1, -1, -1]]
```

Figura 3.16: Algoritmo para cálculo dos momentos de uma distribuição empírica utilizando a função Lauricella.

Deseja-se agora analisar o caso da função Lauricella tal que

$$F_A^{(p)}(s+\alpha(p+1);\alpha,\ldots,\alpha;\alpha+1,\ldots,\alpha+1;-1,\ldots,-1).$$

A Tabela 3.12 apresenta os resultados desse estudo numérico. E, assim, para as parametrizações propostas nos dois casos, observa-se a possibilidade do uso da função de Lauricella no contexto probabilístico.

Tabela 3.12: Cálculo dos momentos de uma distribuição empírica utilizando a função Lauricella para $p = 3,5,8$ e 10, $\gamma_i = \alpha+1$, $x_i = -1$ e $k^{(p)} = s+\alpha(p+1)$ em que $s = 1,2,3$ e 4.

α	s	$k^{(3)}$	$F_A^{(3)}$	$k^{(5)}$	$F_A^{(5)}$	$k^{(8)}$	$F_A^{(8)}$	$k^{(10)}$	$F_A^{(10)}$
0,5	1	3,0	0,19047	4,0	0,04081	5,5	0,00260	6,5	0,00033
	2	4,0	0,12563	5,0	0,02231	6,5	0,00115	7,5	0,00013
	3	5,0	0,08789	6,0	0,01313	7,5	0,00055	8,5	0,00006
	4	6,0	0,06466	7,0	0,00826	8,5	0,00029	9,5	0,00003
1,5	1	7,0	0,00205	10,0	$5{,}52\times10^{-6}$	14,5	$1{,}54\times10^{-10}$	17,5	$6{,}34\times10^{-14}$
	2	8,0	0,00102	11,0	$2{,}12\times10^{-6}$	15,5	$4{,}47\times10^{-11}$	18,5	$1{,}60\times10^{-14}$
	3	9,0	0,00054	12,0	$8{,}65\times10^{-7}$	16,5	$1{,}39\times10^{-11}$	19,5	$4{,}27\times10^{-15}$
	4	10,0	0,00030	13,0	$3{,}75\times10^{-7}$	17,5	$4{,}56\times10^{-12}$	20,5	$1{,}22\times10^{-15}$
2,5	1	11,0	$1{,}42\times10^{-5}$	16,0	$3{,}26\times10^{-10}$	23,5	$2{,}16\times10^{-18}$	28,5	$1{,}91\times10^{-24}$
	2	12,0	$6{,}27\times10^{-6}$	17,0	$1{,}08\times10^{-10}$	24,5	$5{,}27\times10^{-19}$	29,5	$4{,}00\times10^{-25}$
	3	13,0	$2{,}89\times10^{-6}$	18,0	$3{,}74\times10^{-11}$	25,5	$1{,}35\times10^{-19}$	30,5	$8{,}79\times10^{-26}$
	4	14,0	$1{,}39\times10^{-6}$	19,0	$1{,}36\times10^{-11}$	26,5	$3{,}65\times10^{-20}$	31,5	$2{,}03\times10^{-26}$

3.2.3 Estudo Numérico considerando a função generalizada de Kampé de Fériet

Faz-se aqui o estudo numérico utilizando a função de Kampé de Fériet com o intuito de, também, generalizar o uso da função hipergeométrica quando aplicada às distribuições beta generalizadas. Portanto, calcula-se os momentos das distribuições beta beta e beta Student e, assim, verifica-se a possibilidade de utilizar a função generalizada de Kampé de

Fériet para tal fim.

Mediante isto, utiliza-se a função generalizada de Kampé de Fériet cuja expressão é dada por

$$F_{1:1}^{1:2}((a):(c_1,d_1);\ldots;(c_n,d_n);(c):(a+b);(f_1);\ldots;(d_n);s_1,\ldots,s_n)$$
$$=\frac{1}{B(a,b)}\int_0^1 x^{a-1}(1-x)^{b-1}\ {}_2F_1(c_1,d_1;f_1;s_1x)\cdots\ {}_2F_1(c_n,d_n;f_n;s_nx)dx$$

(EXTON, 1978, equação 2.1.5.15). Se $n=2$, então essas funções se reduzem as funções hipergeométricas de Appell F_2, F_3, F_4 e F_1, respectivamente. Se $n=1$, todas as quatro se tornam a função hipergeométrica de Gauss ${}_2F_1$ (EXTON, 1978, p. 29).

Considere-se, inicialmente, a f.d.p. da distribuição beta beta dada pela equação (3.11). Dessa forma, utiliza-se a função generalizada de Kampé de Fériet para estudar o caso

$${}_2F_1^{(p)}((s+\alpha(p+1)):(1-\beta,\alpha);\ldots;(1-\beta,\alpha);(\beta+s+\alpha(p+1)):(\alpha+1),\ldots,(\alpha+1);1,\ldots,1),$$

em que se calcula os momentos da distribuição beta beta. A Tabela 3.13 apresenta o cálculo desses momentos quando usada a função generalizada de Kampé de Fériet.

Tabela 3.13: Cálculo dos momentos da beta beta utilizando a função generalizada de Kampé de Fériet para $p=1,2,5$ e 10 e $s=1,2,3$ e 4.

$(\alpha;\beta)$	s	${}_2F_1^{(1)}$	${}_2F_1^{(2)}$	${}_2F_1^{(5)}$	${}_2F_1^{(10)}$
$(1,0;1,5)$	1	0,80208	0,61427	0,24117	0,04199
	2	0,78125	0,59041	0,22862	0,03973
	3	0,76628	0,57266	0,21865	0,03784
	4	0,75495	0,55890	0,21056	0,03624
$(2,5;1,0)$	1	1,00000	1,00000	1,00000	1,00000
	2	1,00000	1,00000	1,00000	1,00000
	3	1,00000	1,00000	1,00000	1,00000
	4	1,00000	1,00000	1,00000	1,00000
$(1,0;2,0)$	1	0,70000	0,45238	0,09610	0,00513
	2	0,66667	0,41964	0,08641	0,00457
	3	0,64286	0,39583	0,07907	0,00413
	4	0,62500	0,37778	0,07336	0,00378
$(3,5;2,0)$	1	0,37778	0,11913	0,00248	$2{,}43\times10^{-6}$
	2	0,36364	0,11321	0,00234	$2{,}30\times10^{-6}$
	3	0,35185	0,10818	0,00222	$2{,}18\times10^{-6}$
	4	0,34188	0,10385	0,00211	$2{,}07\times10^{-6}$

Segue-se, na Figura 3.17, uma exemplificação da utilização do algoritmo para cálculo dos momentos da distribuição beta beta no caso em que $p=1$, $s=4$, $\alpha=3,5$ e $\beta=2,0$. Este algoritmo foi implementado utilizando o software MATHEMATICA® como ferramenta.

```
In[1]:= FB = Function[{a, b, alpha, bta, x1},
          (Beta[a, b])^-1 * ∫_0^1 t^(a-1) * (1-t)^(b-1) * Hypergeometric2F1[1-bta, alpha, alpha+1, x1*t]
            dt];

In[2]:= Clear[a]; Clear[s]; Clear[p]; Clear[alpha]; Clear[b];
        p = 1;
        s = 4;
        alpha = 3.5;
        bta = 2.0;
        a = s + alpha*(p+1);
        b = bta + s + alpha*(p+1) - a;
        N[FB[s + alpha*(p+1), b, alpha, bta, 1]]
```

Figura 3.17: Algoritmo referente à obtenção dos momentos para a distribuição beta beta utilizando a função generalizada de Kampé de Fériet.

Considere-se, agora, a f.d.p. da distribuição beta Student $BS(a,b,\nu)$ cuja expressão é apresentada pela equação (3.10). Para cálculo dos momentos da distribuição beta Student, faz-se uso da função generalizada de Kampé de Fériet de forma que sua função seja

$$ {}_2F_1^{(p)}\left(\left(\frac{s+p+1}{2}\right):\left(1-\frac{\nu}{2},\frac{1}{2}\right);\ldots;\left(1-\frac{\nu}{2},\frac{1}{2}\right);\left(\frac{\nu+p+1}{2}\right):\left(\frac{3}{2}\right),\ldots,\left(\frac{3}{2}\right);1,\ldots,1\right). \tag{3.12} $$

A Tabela 3.14 apresenta o resultado numérico dos momentos dessa distribuição para diferentes valores de p e s quando $\nu = 6$ e 8. O algoritmo apresentado na Figura 3.17 pode, também, ser utilizado para cálculo dos momentos da distribuição beta Student t, sendo preciso apenas modificar as parametrizações conforme estrutura contida na equação 3.12.

Tabela 3.14: Cálculo dos momentos da beta Student utilizando a função generalizada de Kampé de Fériet para $p = 2,4,6$ e 8 e $s = 1,2,3$ e 4.

ν	s	${}_2F_1^{(2)}$	${}_2F_1^{(4)}$	${}_2F_1^{(6)}$	${}_2F_1^{(8)}$
6	1	0,57632	0,26951	0,06666	0,00331
	2	0,49971	0,21274	0,03043	0,00123
	3	0,43302	0,16728	0,01204	0,00040
	4	0,37541	0,13121	0,00383	0,00010
8	1	0,53772	0,22398	0,07667	0,00273
	2	0,46235	0,17424	0,03703	0,00107
	3	0,39894	0,13590	0,01628	0,00038
	4	0,34582	0,10646	0,00637	0,00012

3.3 Considerações Finais

Neste Capítulo apresentou-se a distribuição beta e a generalização desta, conhecida na literatura como distibuição beta generalizada. Realizou-se, por conseguinte, uma revisão literária referente a algumas distribuições beta generalizadas e suas funcionalidades. De modo ilustrativo, obteve-se os momentos ordinários destas distribuições mediante análise numérica e, ainda, os gráficos de assimetria e curtose. Um ponto importante deste capítulo pode ser visto nas Seções 3.2.2 e 3.2.3 referentes às funções de Lauricella e generalizada de Kampé de Fériet, pois mostra como obter os momentos de maneira mais objetiva. Para o caso da função generalizada de Kampé de Fériet, foram calculados os momentos das distribuições beta beta e beta Student.

4 A Distribuição Beta Power

Propõe-se uma generalização da distribuição power, a qual será chamada de distribuição *beta power* (BP). Encontrar-se-á a forma analítica da f.d.p. correspondente e a função da razão de risco, sendo estas ilustradas através de gráficos. Calcula-se o n-ésimo momento, o momento das estatísticas de ordem e, ainda, analisa-se a variação das medidas de assimetria e curtose. Discorre-se, também, a estimação dos parâmetros através do método de máxima verossimilhança. Além disto, ilustra-se a utilidade da distribuição beta power através de aplicações a dados reais e a dados simulados.

4.1 Distribuição Beta Power

A distribuição BP é apresentada seguindo a idéia proposta por Eugene et al. (2002) e cuja correspondente definição foi descrita na Seção 3.2. Portanto, para obtenção da distribuição BP é sabido da necessidade de se conhecer a f.d.p. e a f.d.a. da distribuição power tal que estas são, respectivamente, expressas como

$$g(x) = \alpha\beta^{\alpha}x^{\alpha-1}, \quad 0 < x < \frac{1}{\beta}, \quad \alpha > 0, \quad \beta > 0 \tag{4.1}$$

e

$$G(x) = (\beta x)^{\alpha}, \tag{4.2}$$

em que α é o parâmetro de forma e β é o parâmetro de escala. No caso especial, quando $\alpha = 1$, tem-se uma distribuição retangular no intervalo $[0, 1/\beta]$ que implica na distribuição power $P(1, \beta)$ cuja distribuição é equivalente a $X \sim U(0, 1/\beta)$. Em Balakrishnan e Nevzorov (2003) é apresentada a distribuição power padrão, ou seja, quando $\beta = 1$, de forma a $g(x) = \alpha x^{\alpha-1}$ e $G(x) = x^{\alpha}$.

Através da equação (3.4) e substituindo $G(x)$ pela f.d.a. (4.2) da distribuição power, obtém-se a f.d.a. da distribuição BP

$$F(x) = I_{(\beta x)^{\alpha}}(a, b) \tag{4.3}$$

para $0 < x < 1/\beta$, $a > 0$, $b > 0$, $\alpha > 0$ e $\beta > 0$. Para a e b gerais, pode-se expressar (4.3) em termos da bem conhecida função hipergeométrica expressa por

$$F(x) = \frac{\{(\beta x)^\alpha\}^a}{aB(a,b)}\, {}_2F_1(a, 1-b, a+1; (\beta x)^\alpha).$$

A f.d.p. $f(x)$ e a função da razão de risco $h(x)$ associada a (4.3) são

$$f(x) = \frac{\alpha\beta(\beta x)^{\alpha a-1}\{1-(\beta x)^\alpha\}^{b-1}}{B(a,b)}, \quad 0 < x < \frac{1}{\beta} \tag{4.4}$$

e

$$h(x) = \frac{\alpha\beta(\beta x)^{\alpha a-1}\{1-(\beta x)^\alpha\}^{b-1}}{B(a,b)\{1-I_{(\beta x)^\alpha}(a,b)\}}. \tag{4.5}$$

As Figuras 4.1 e 4.2 ilustram algumas das possíveis formas da f.d.p. (4.4) e da função da razão de risco (4.5), respectivamente, para valores selecionados de parâmetros, incluindo o caso da distribuição power.

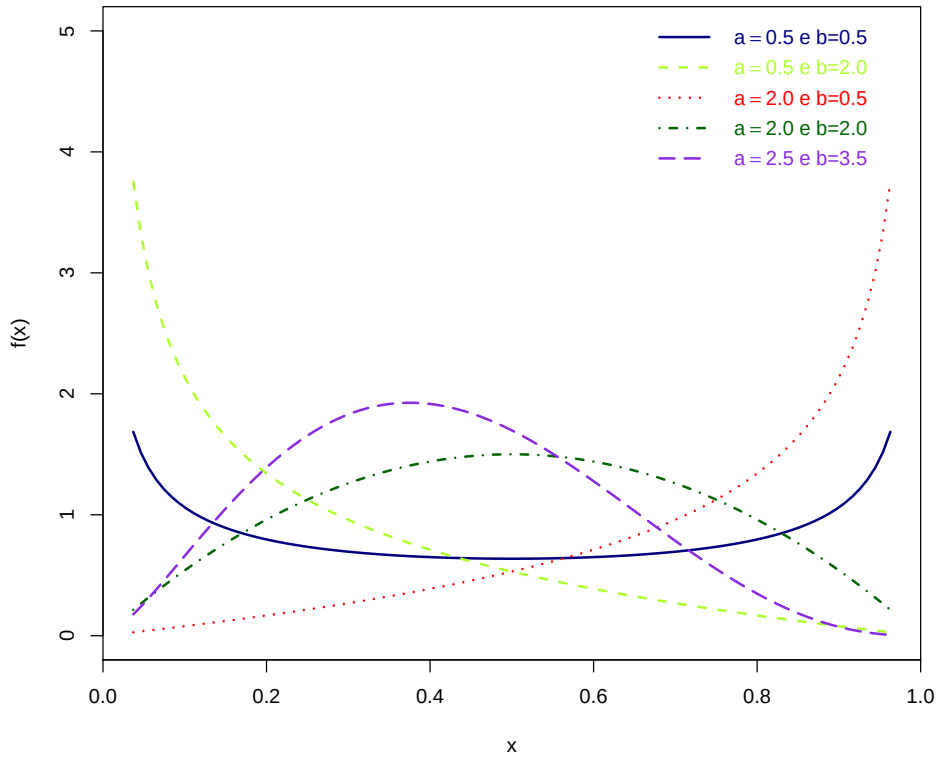

Figura 4.1: Gráfico da f.d.p. da distribuição $BP(a,b,1,1)$ para valores selecionados de parâmetros.

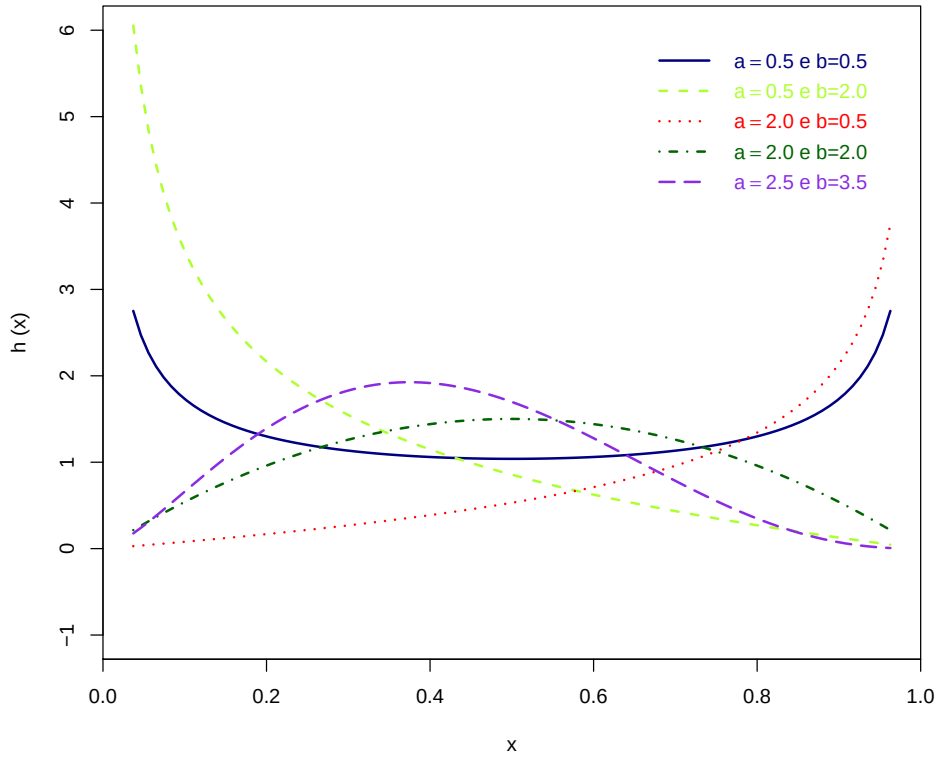

Figura 4.2: Gráfico da $h(x)$ da distribuição $BP(a,b,1,1)$ para valores selecionados de parâmetros.

4.2 Expansão das Funções de Densidade e de Distribuição

A f.d.p. em (4.4) é de fácil implementação em qualquer software estatístico. Aqui, obtém-se simples expressões da f.d.a. da distribuição BP que depende dos parâmetros b (ou a) ser inteiro ou real não-inteiro. Considera-se a expansão em série

$$(1-z)^{b-1} = \sum_{i=0}^{\infty} \frac{(-1)^i \Gamma(b)}{\Gamma(b-i)i!} z^i, \tag{4.6}$$

válida para $|z| < 1$ e $b > 0$ real não-inteiro. Aplicações de (4.6) em (4.3) se b é real não-inteiro é expressa como

$$F(x) = \frac{\Gamma(a+b)}{\Gamma(a)} \sum_{i=0}^{\infty} \frac{(-1)^i G(x)^{a+i}}{\Gamma(b-i)i!(a+i)}, \tag{4.7}$$

em que $G(x)$ provém de (4.2). Dessa maneira, tem-se

$$F(x) = \frac{\Gamma(a+b)}{\Gamma(a)} \sum_{i=0}^{\infty} \frac{(-1)^i (\beta x)^{\alpha(a+i)}}{\Gamma(b-i) i! (a+i)}. \tag{4.8}$$

Para b inteiro, a soma em (4.8) pára em $b-1$.

4.3 Momentos

Se X tem f.d.p. (4.4), então seu n-ésimo momento pode ser escrito como

$$E(X^n) = \frac{1}{B(a,b)} \int_0^{1/\beta} x^n \left[\alpha\beta (\beta x)^{\alpha a - 1} \{1 - (\beta x)^{\alpha}\}^{b-1} \right] dx. \tag{4.9}$$

Usando a expansão binomial de (4.6), (4.9) pode ser re-escrita como

$$E(X^n) = \frac{\Gamma(a+b)}{\Gamma(a)} \sum_{j=0}^{\infty} \frac{(-1)^j I(k)}{\Gamma(b-j) j!}, \tag{4.10}$$

em que $I(k)$ denota a integral

$$I(k) = \int_0^{1/\beta} \alpha x^{n-1} G(x)^{a+j} dx = \alpha [\beta^n \{n + \alpha(a+j)\}]^{-1}. \tag{4.11}$$

Finalmente, (4.9) pode ser calculada como

$$E(X^n) = \frac{\alpha \Gamma(a+b)}{\beta^n \Gamma(a)} \sum_{j=0}^{\infty} (-1)^j \{\Gamma(b-j)\{n + \alpha(a+j)\} j!\}^{-1}. \tag{4.12}$$

É possível obter, ainda, a forma fechada do n-ésimo momento da distribuição beta power. Sendo assim, tem-se que esse momento é expresso como

$$E(X^n) = \frac{B(a + \frac{n}{\alpha}, b)}{\beta^n B(a,b)}. \tag{4.13}$$

As Tabelas 4.1 e 4.2 apresentam os momentos ordinários numericamente para a distribuição $BP(a,b,1,1)$. As Figuras 4.3 e 4.4 ilustram a variação das medidas de assimetria e curtose, as quais são controladas pelos parâmetros a e b e que são calculadas usando (4.9). Ao analisar as curvas de assimetria, tem-se que para a fixo, a assimetria decresce enquanto que, para b fixo, ocorre o contrário. Enquanto na curtose, para a fixo ocorre um decréscimo para valores de b menores que dois. E, para valores de b a partir de dois, a curtose passa a crescer. O mesmo ocorre para o caso em que $BP(a,b,1,1)$ é função de b para valores fixos de a. Nota-se, ainda, que para $b = 1,0$ como função de a e ainda $a = 1,0$

como função de b há um crescimento da curtose em todo o intervalo $[1,6]$.

Tabela 4.1: Momentos ordinários da distribuição BP para diferentes valores de α, β, a e b.

μ'_k	$BP(1{,}0,\ 1{,}0,\ 1,\ 1)$	$BP(1{,}0,\ 1{,}5,\ 1,\ 1)$	$BP(1{,}0,\ 3{,}5,\ 1,\ 1)$
μ'_1	0,50000	0,40000	0,22222
μ'_2	0,33333	0,22857	0,08081
μ'_3	0,25000	0,15238	0,03730
μ'_4	0,20000	0,11082	0,01989
μ'_5	0,16667	0,08525	0,01170
μ'_6	0,14286	0,06820	0,00739
σ^2	0,08333	0,06857	0,03142
α_3	0	0,33945	0,96428
α_4	1,80000	2,05050	3,40880

Tabela 4.2: Momentos ordinários da distribuição BP para diferentes valores de α, β, a e b.

μ'_k	$BP(1{,}5,\ 1{,}5,\ 1,\ 1)$	$BP(1{,}5,\ 2{,}5,\ 1,\ 1)$	$BP(2{,}5,\ 3{,}5,\ 1,\ 1)$
μ'_1	0,50000	0,375	0,37500
μ'_2	0,31250	0,1875	0,18750
μ'_3	0,21875	0,10938	0,10938
μ'_4	0,16406	0,070313	0,07031
μ'_5	0,12891	0,04834	0,04834
μ'_6	0,10474	0,034912	0,03491
σ^2	0,06250	0,046875	0,04687
α_3	0	0,3849	0,38490
α_4	2,00000	2,3333	2,33330

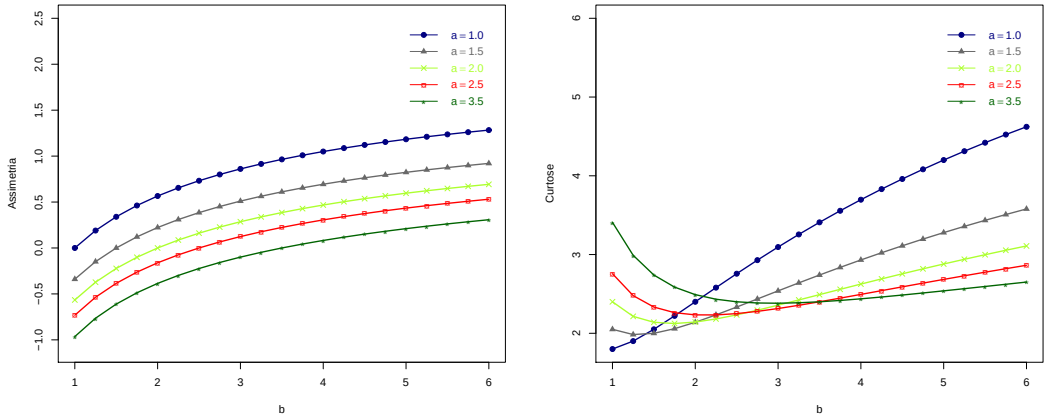

Figura 4.3: Gráficos de assimetria e curtose para a distribuição $BP(a,b,1,1)$ como função de b e fixado a.

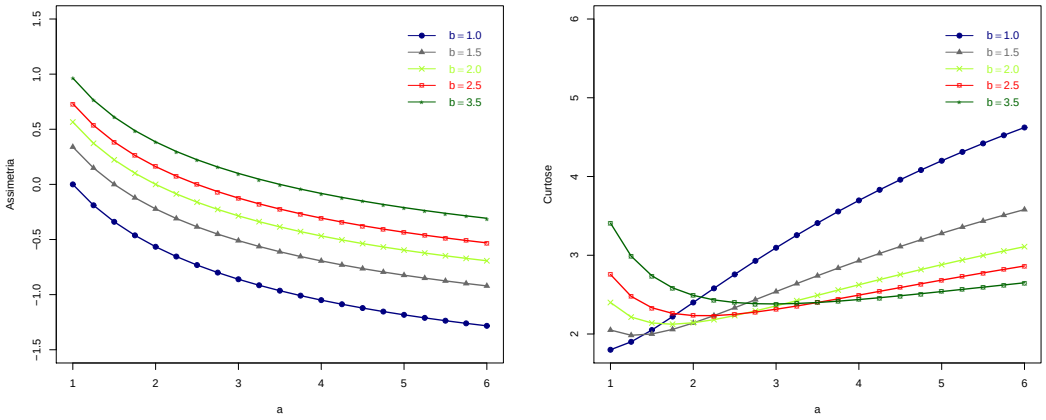

Figura 4.4: Gráficos de assimetria e curtose para a distribuição $BP(a,b,1,1)$ como função de a e fixado b.

4.4 Estatísticas de Ordem

A função densidade da i-ésima estatística de ordem $X_{i:n}$, dita $f_{i:n}(x)$, numa amostra aleatória de tamanho n da distribuição BP é conhecida como

$$f_{i:n}(x) = \frac{f(x)}{B(i,n-i+1)} F(x)^{i-1}\{1-F(x)\}^{n-i},$$

para $i = 1,\ldots,n$.

Usando (4.3) e (4.4), pode ser expressa $f_{i:n}(x)$ em termos da razão da função beta incompleta

$$f_{i:n}(x) = \frac{\alpha\beta(\beta x)^{\alpha a-1}\{1-(\beta x)^{\alpha}\}^{b-1}}{B(a,b)B(i,n-i+1)} I_{(\beta x)^{\alpha}}(a,b)^{i-1} I_{1-(\beta x)^{\alpha}}(a,b)^{n-i}$$

ou ainda em termos da função hipergeométrica como

$$f_{i:n}(x) = \frac{\alpha\beta(\beta x)^{\alpha ai-1}\{1-(\beta x)^{\alpha}\}^{a(n-1)+b-1}}{a^{n-1}B(a,b)^{n}B(i,n-i+1)} \; {}_2F_1(a,1-b,a+1;(\beta x)^{\alpha})^{i-1}$$
$$\times \; {}_2F_1(a,1-b,a+1;1-(\beta x)^{\alpha})^{n-i}.$$

Os momentos das estatísticas de ordem podem ser obtidos diretamente dos momentos da distribuição BP. Ou, ainda, podem ser obtidos usando o seguinte resultado, devido a Barakat e Abdelkander (2004), aplicado ao caso em que a variável é independente e identicamente distribuída

$$E(X_{i:n}^k) = k \sum_{j=n-i+1}^{n} (-1)^{j-n+i-1} \binom{j-1}{n-i} \binom{n}{j} I_j(k), \tag{4.14}$$

em que

$$I_j(k) = \int_0^{1/\beta} x^{k-1}\{1-F(x)\}^j dx.$$

Sabe-se que $(y+x)^d = \binom{d}{0}y^d + \binom{d}{1}y^{d-1} + \binom{d}{2}y^{d-2}x^2 + \ldots = \sum_{l=0}^{d}\binom{d}{l}y^{d-l}x^l$.

Para $\{1-F(x)\}^j$, tem-se que

$$\begin{aligned}\{1-F(x)\}^j &= \sum_{l=0}^{j}\binom{j}{l}\{-F(x)\}^l = \sum_{l=0}^{j}(-1)^l\binom{j}{l}F(x)^l \\ &= \sum_{l=0}^{j}(-1)^l\binom{j}{l}\left\{\frac{\Gamma(a+b)}{\Gamma(a)}\sum_{i=0}^{\infty}\frac{(-1)^i(\beta x)^{\alpha(a+i)}}{\Gamma(b-i)i!(a+i)}\right\}^l.\end{aligned}$$

Logo,

$$I_j(k) = \int_0^{1/\beta} x^{k-1}\sum_{l=0}^{j}(-1)^l\binom{j}{l}\left\{\frac{\Gamma(a+b)}{\Gamma(a)}\right\}^l\left\{\sum_{i=0}^{\infty}\frac{(-1)^i(\beta x)^{\alpha(a+i)}}{\Gamma(b-i)i!(a+i)}\right\}^l.$$

Seja, ainda, $(\sum_{i=0}^{\infty}a_i)^z = \sum_{[m_1,\ldots,m_z]=0}^{\infty}a_{m_1}\cdots a_{m_z}$, para z inteiro positivo. Assim,

$$I_j(k) = \int_0^{1/\beta} x^{k-1}\sum_{l=0}^{j}(-1)^l\binom{j}{l}\left\{\frac{\Gamma(a+b)}{\Gamma(a)}\right\}^l\sum_{[m_1,\ldots,m_l]=0}^{\infty}a_{m_1}\cdots a_{m_l}x^{\alpha\sum_{q=1}^{l}(a+m_q)}, \tag{4.15}$$

em que

$$a_{m_q} = \frac{(-1)^{m_q}\beta^{\alpha(a+m_q)}}{\Gamma(b-m_q)m_q!(a+m_q)}, \quad q = 1,\ldots,l.$$

Deste modo, ao substituir (4.15) em (4.14), obtém-se os momentos das estatísticas de ordem cuja expressão é dada por

$$\begin{aligned}E(X_{i:n}^k) =& k\sum_{j=n-i+1}^{n}\sum_{l=0}^{j}\sum_{[m_1,\ldots,m_l]=0}^{\infty}(-1)^{j-n+i+l-1}\binom{j-1}{n-i}\binom{n}{j}\binom{j}{l}\left\{\frac{\Gamma(a+b)}{\Gamma(a)}\right\}^l \\ &\times a_{m_1}\cdots a_{m_l}\left[\beta^{k+\alpha\sum_{q=1}^{l}(a+m_q)}\{k+\alpha\sum_{q=1}^{l}(a+m_q)\}\right]^{-1}.\end{aligned}$$

4.5 Estimação

Considera-se que X_i segue a distribuição BP e seja $\theta = (a,b,\alpha,\beta)^T$ como sendo o vetor de parâmetros. O logaritmo da função de verossimilhança $\ell = \ell(\theta)$ para θ de uma amostra aleatória $x_1,\ldots,x_n$ de (4.4) é expressa como

$$\ell = n\log(\alpha\beta) - n\log B(a,b) + (\alpha a - 1)\sum_{i=1}^{n}\log(\beta x_i) + (b-1)\sum_{i=1}^{n}\log\{1-(\beta x_i)^\alpha\}.$$

Os componentes do vetor escore $U = U(\theta) = (\partial\ell/\partial a, \partial\ell/\partial b, \partial\ell/\partial\alpha, \partial\ell/\partial\beta)^T$ para n observações são apresentados como

$$\begin{aligned}
\frac{\partial\ell}{\partial a} &= -n\psi(a) + n\psi(a+b) + \alpha\sum_{i=1}^{n} log(\beta x_i), \\
\frac{\partial\ell}{\partial b} &= -n\psi(b) + n\psi(a+b) + \sum_{i=1}^{n}\log\{1-(\beta x_i)^{\alpha}\}, \\
\frac{\partial\ell}{\partial\alpha} &= \frac{n}{\alpha} + a\sum_{i=1}^{n}\log(\beta x_i) - (b-1)\sum_{i=1}^{n}\frac{(\beta x_i)^{\alpha}\log(\beta x_i)}{1-(\beta x_i)^{\alpha}}, \\
\frac{\partial\ell}{\partial\beta} &= \frac{\alpha}{\beta}\Big[na - (b-1)\sum_{i=1}^{n}\frac{(\beta x_i)^{\alpha}}{1-(\beta x_i)^{\alpha}}\Big].
\end{aligned}$$

em que $\psi(x) = \partial\log\Gamma(x)/\partial x$ é a função *digamma*. Pode-se obter as estimativas de máxima-verossimilhança dos quatro parâmetros quando suas expressões são igualadas a zero e resolvendo-as simultaneamente. A obtenção dessas estimativas pode ser realizada através, por exemplo, do método escore de Fisher ou método de Newton-Raphson pois o suporte da distribuição beta power depende do parâmetro.

4.6 Aplicações

Com o intuito de ilustrar a funcionalidade da distribuição BP, foram analisados quatro conjunto de dados sendo dois referentes a dados reais e os demais correspondentes a dados simulados. Para os quatro exemplos aqui abordados, pôde ser observada uma boa performance da distribuição BP.

Primeiro Conjunto de Dados Reais: Os dados a serem analisados são provenientes de medidas de amostras de rochas de petróleo, consistindo de 48 amostras de rochas de um reservatório de petróleo. Esse conjunto de dados corresponde a 12 amostras de núcleo de um reservatório de petróleo que foram amostradas por 4 seções cruzadas. Cada amostra de núcleo foi medida para permeabilidade e cada seção cruzada apresentou o total de área dos poros, total do perímetro dos poros e a forma. Neste estudo referente ao ajuste da distribuição BP, considera-se a análise da variável forma do perímetro pela área.

Para obtenção das estimativas, os chutes iniciais assumiram os seguintes valores: $a = 55,0$, $b = 98,0$, $\alpha = 0,4$, $\beta = 0,2$. Os parâmetros estimados para a distribuição beta power foram: $\hat{\alpha} = 0,2949$, $\hat{\beta} = 0,1561$, $\hat{a} = 56,0247$ e $\hat{b} = 97,8101$ cujo $AIC = -106,5240$. E, em se tratando da distribuição power, os parâmetros estimados foram $\hat{\alpha} = 1,1506$, $\hat{\beta} = 2,1546$

e $AIC = -71,3068$. Ao comparar as distribuições pelo critério de informação de Akaike (AIC), observa-se que a distribuição beta power apresenta o menor valor de AIC.

Tabela 4.3: Forma do perímetro pela aréa correspondente ao conjunto de dados de medidas da amostra de rochas de petróleo.

0,0903296	0,2036540	0,2043140	0,2808870	0,1976530	0,3286410
0,1486220	0,1623940	0,2627270	0,1794550	0,3266350	0,2300810
0,1833120	0,1509440	0,2000710	0,1918020	0,1541920	0,4641250
0,1170630	0,1481410	0,1448100	0,1330830	0,2760160	0,4204770
0,1224170	0,2285950	0,1138520	0,2252140	0,1769690	0,2007440
0,1670450	0,2316230	0,2910290	0,3412730	0,4387120	0,2626510
0,1896510	0,1725670	0,2400770	0,3116460	0,1635860	0,1824530
0,1641270	0,1534810	0,1618650	0,2760160	0,2538320	0,2004470

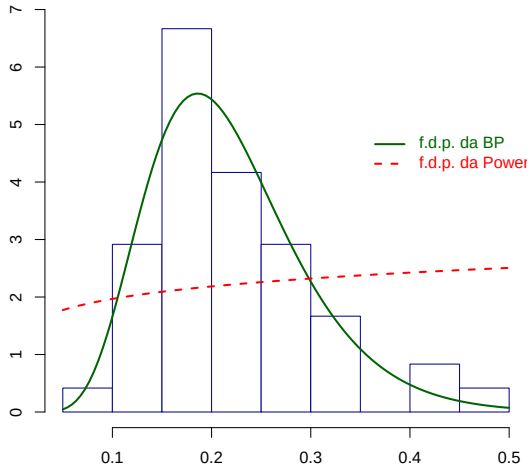

Figura 4.5: Histograma do primeiro conjunto de dados reais e as correspondentes f.d.p. para as distribuições BP e power.

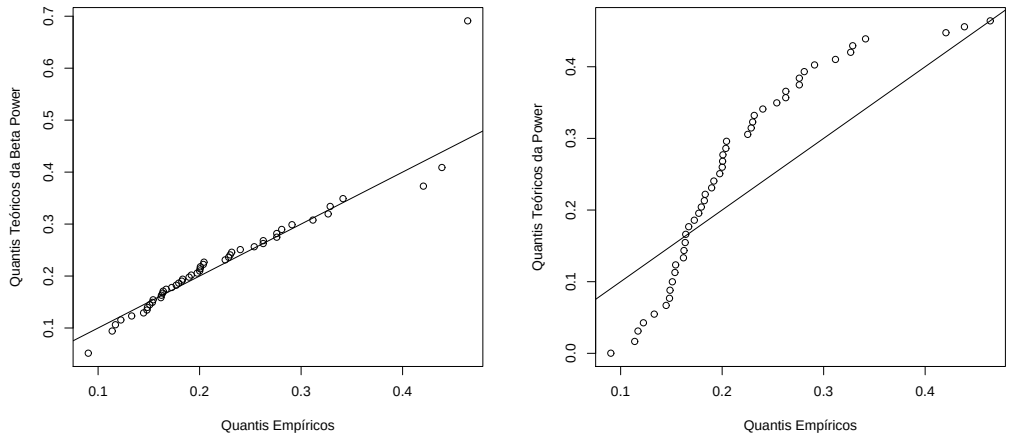

Figura 4.6: Gráficos dos quantis teóricos versus quantis empíricos para as distribuições BP e power referentes ao primeiro conjunto de dados reais.

A Figura 4.5 apresenta o histograma do correspondente conjunto de dados juntamente com as densidades da distribuição beta power e distribuição power relacionadas às respectivas estimativas de parâmetros. Na referente figura é possível visualizar a boa performance da distribuição beta power. Através da Figura 4.6, a qual apresenta os gráficos de quantis teóricos versus quantis empíricos para as distribuições BP e power, pode ser observado que a distribuição beta power apresenta-se adequada ao conjunto de dados pois os pontos correspondentes ao conjunto de dados mostram-se próximos à reta de 45 graus.

Segundo Conjunto de Dados Reais: Ajusta-se a distribuição BP aos dados referentes a produção total de leite (*PLTOTAL*) de vacas da raça SINDI conforme descrito na Seção 3.1.1. Considerando apenas a variável *PLTOTAL* para ser ajustada, faz-se uma breve transformação tal que esta corresponda a dados de proporção, ou seja, $y = (PLTOTAL - min)/(max - min)$. Os dados estão apresentados na Tabela 4.4.

Os chutes iniciais foram: $a = 0,5$, $b = 42,0$, $\alpha = 4,0$, $\beta = 0,8$. Por conseguinte, os parâmetros estimados referentes à distribuição BP apresentam aos seguintes valores: $\hat{\alpha} = 6,6402$, $\hat{\beta} = 0,7756$, $\hat{a} = 0,2704$ e $\hat{b} = 42,0228$ cujo $AIC = -47,9976$. Para a distribuição power, tem-se que os parâmetros estimados para a variável correspondente foram: $\hat{\alpha} = 5,0027$ e $\hat{\beta} = 1,1364$, e cujo $AIC = 289,6174$. Nota-se, através dos valores de AIC que a distribuição BP mostra-se satisfatória.

Tabela 4.4: Produção total de leite em proporção correspondente ao conjunto de dados apresentado na Tabela 3.1.

0,4365	0,4260	0,5140	0,6907	0,7471	0,2605	0,6196
0,8781	0,4990	0,6058	0,6891	0,5770	0,5394	0,1479
0,2356	0,6012	0,1525	0,5483	0,6927	0,7261	0,3323
0,0671	0,2361	0,4800	0,5707	0,7131	0,5853	0,6768
0,5350	0,4151	0,6789	0,4576	0,3259	0,2303	0,7687
0,4371	0,3383	0,6114	0,3480	0,4564	0,7804	0,3406
0,4823	0,5912	0,5744	0,5481	0,1131	0,7290	0,0168
0,5529	0,4530	0,3891	0,4752	0,3134	0,3175	0,1167
0,6750	0,5113	0,5447	0,4143	0,5627	0,5150	0,0776
0,3945	0,4553	0,4470	0,5285	0,5232	0,6465	0,0650
0,8492	0,8147	0,3627	0,3906	0,4438	0,4612	0,3188
0,2160	0,6707	0,6220	0,5629	0,4675	0,6844	-
0,3413	0,4332	0,0854	0,3821	0,4694	0,3635	-
0,4111	0,5349	0,3751	0,1546	0,4517	0,2681	-
0,4049	0,5553	0,5878	0,4741	0,3598	0,7629	-
0,5941	0,6174	0,6860	0,0609	0,6488	0,2747	-

A Figura 4.7 apresenta o histograma do correspondente conjunto de dados, tal que pode ser verificada a boa performance da distribuição beta power, juntamente com as densidades da distribuição beta power e distribuição power relacionadas com as respectivas

estimativas de parâmetros. Observa-se, também, mediante Figura 4.8, que a distribuição BP se mostra adequada ao conjunto de dados.

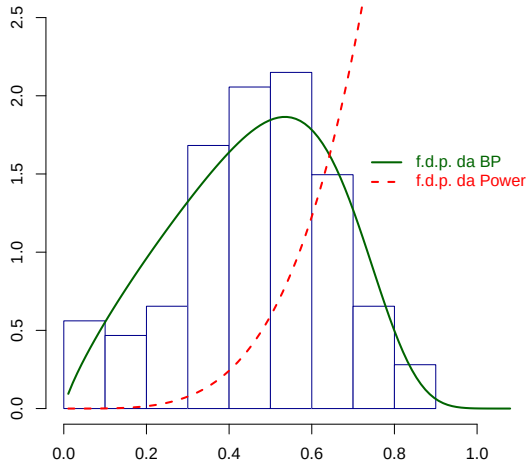

Figura 4.7: Histograma do segundo conjunto de dados reais e as correspondentes f.d.p. para as distribuições BP e power.

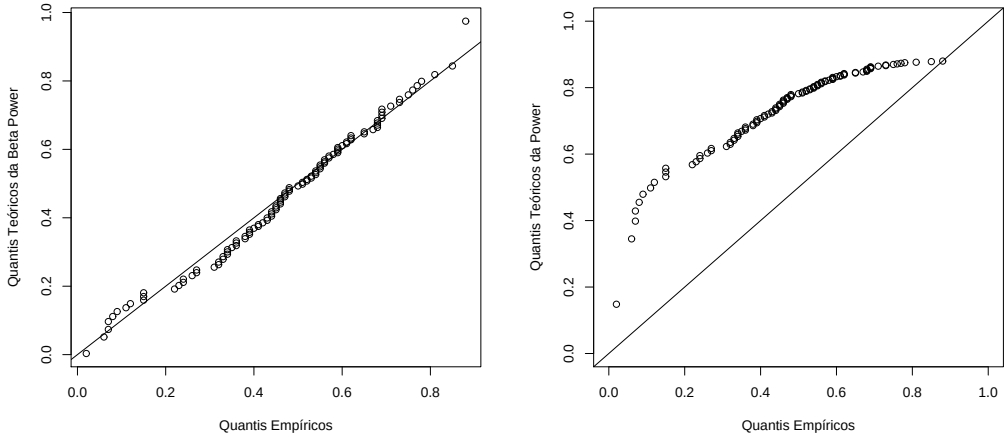

Figura 4.8: Gráficos dos quantis teóricos versus quantis empíricos para as distribuições BP e power referentes ao segundo conjunto de dados reais.

Primeiro Conjunto de Dados Simulados: Utilizando o gerador de números aleatórios uniformes do Marsaglia, foram gerados 100 números aleatórios tendo como base a distribuição power $P(0,5;\ 11)$. A obtenção dos números aleatórios ocorre mediante uso do algotimo apresentado na Figura 4.9.

Após geração dos números aleatórios, propõe-se os seguintes chutes iniciais: $a = 9,0$, $b = 1,0$, $\alpha = 0,5$ e $\beta = 11,0$. Por conseguinte, as estimativas obtidas para os parâmetros da distribuição beta power são: $\hat{\alpha} = 0,0663$, $\hat{\beta} = 11,1970$, $\hat{a} = 7,2638$ e $\hat{b} = 0,8875$.

```
RNGkind("Marsaglia-Multicarry")
x<-c()
alpha<- 0.5
beta<- 11
x<-(runif(100)^(1/alpha))/beta
x
```

Figura 4.9: Algoritmo para obtenção de números aleatórios da distribuição power

A Figura 4.10 apresenta o histograma do correspondente conjunto de dados simulados juntamente com a função densidade de probabilidade das distribuições BP e power. Com relação à distribuição power, os parâmetros estimados corresponderam a $\hat{\alpha} = 0,5310$ e $\hat{\beta} = 11,1970$.

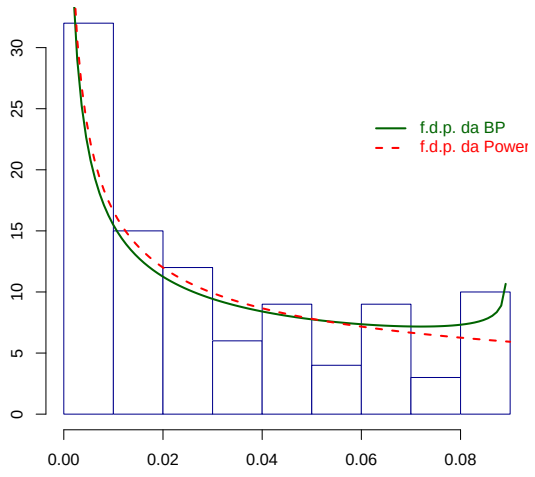

Figura 4.10: Histograma do primeiro conjunto de dados simulados e as correspondentes f.d.p. para as distribuições BP e power.

Neste caso, mediante a Figura 4.10, observa-se um bom desempenho de ambas distribuições. Tem-se, ainda, que para a distribuição beta power o AIC foi de $-533,6194$ e para a distribuição power resultou-se um AIC de $-529,3844$. Através da Figura 4.11, é possível visualizar os gráficos dos quantis teóricos versus os quantis empíricos, entretanto, não é trivial identificar qual distribuição apresenta melhor performance. Dessa forma, ao comparar ambas distribuições pelo critério de informação de Akaike (AIC), observa-se que a distribuição BP apresenta-se melhor adequada.

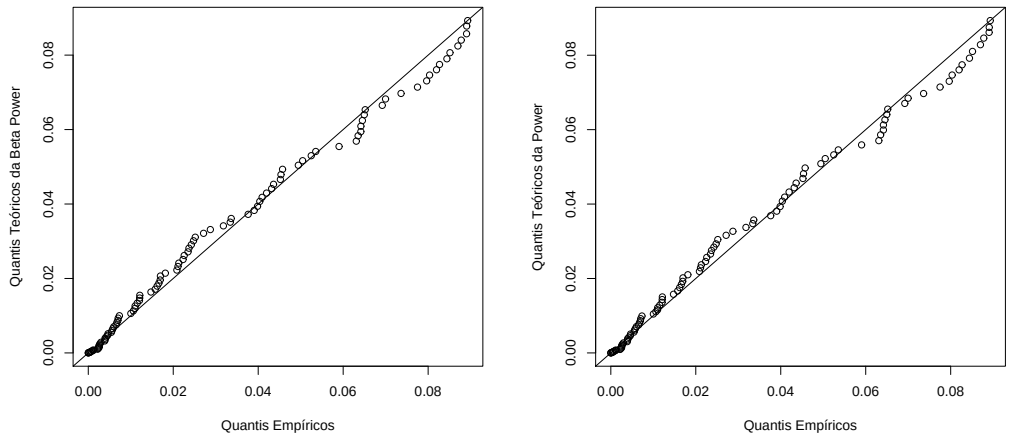

Figura 4.11: Gráficos dos quantis teóricos versus quantis empíricos para as distribuições BP e power referentes ao primeiro conjunto de dados simulados.

Segundo Conjunto de Dados Simulados: Utilizando o gerador de números aleatórios uniformes do Marsaglia, foram gerados 100 números aleatórios, de modo análogo ao exemplo anterior, tendo como base a distribuição power $P(2;2)$. Os chutes iniciais para obter as estimativas são dados por $a = 10,0$, $b = 1,1$, $\alpha = 2,0$ $\beta = 2,0$. As estimativas dos parâmetros para a distribuição beta power foram, portanto, $\hat{\alpha} = 0,1674$, $\hat{\beta} = 2,0008$, $\hat{a} = 10,5708$ e $\hat{b} = 0,0851$. Com respeito à distribuição power, os parâmetros estimados corresponderam a $\hat{\alpha} = 2,0000$, $\hat{\beta} = 2,0008$.

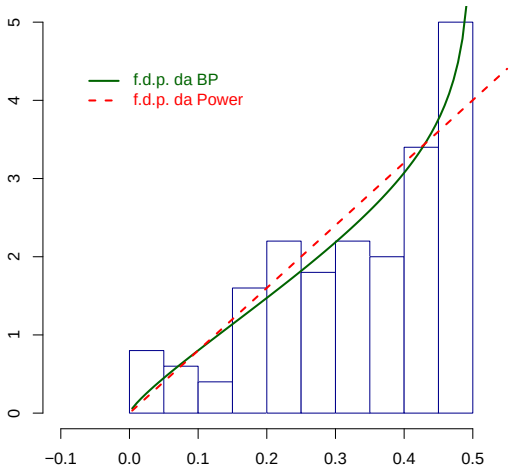

Figura 4.12: Histograma do segundo conjunto de dados simulados e as correspondentes f.d.p. para as distribuições BP e power.

A Figura 4.12 apresenta o histograma do correspondente conjunto de dados simulados juntamente com as f.d.p. referentes às distribuições BP e power, em que se pode observar um bom desempenho de ambas distribuições. A Figura 4.13 apresenta a relação entre

os quantis teóricos e empíricos para as distribuições BP e power, porém, não é possível visualizar alguma diferença notória a partir dos gráficos de modo a identificar qual a distribuição de melhor performance. Sendo assim, ao comparar ambas distribuições pelo AIC, tem-se que para a distribuição beta power o AIC foi de $-173,4301$ e para a distribuição power resultou-se um AIC de $-164,8587$ e, portanto, verifica-se a boa performance da distribuição beta power por apresentar um menor AIC.

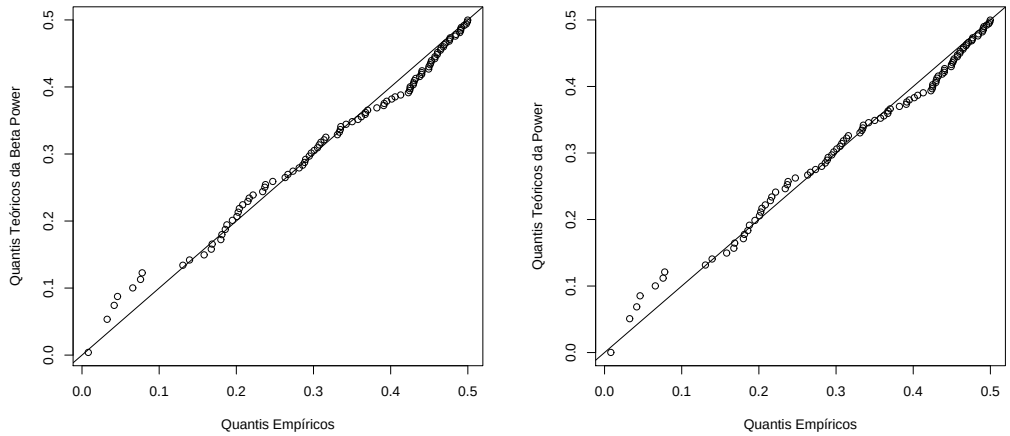

Figura 4.13: Gráficos dos quantis teóricos versus quantis empíricos para as distribuições BP e power referentes ao segundo conjunto de dados simulados.

4.7 Considerações Finais

Ultimamente, pesquisas científicas têm ofertado novas distribuições correspondentes às distribuições beta generalizadas. Conforme visto no Capítulo 3, existem diversas distribuições beta generalizadas as quais apresentam boa performance com respeito às distribuições usuais. Neste Capítulo foi apresentada uma nova distribuição beta generalizada a qual é denominada de distribuição beta power. A distribuição power é pouco conhecida na literatura e, por essa razão, torna-se mais complexo falar da sua aplicabilidade. Portanto, em pesquisas futuras, pretende-se abordar melhor esta distribuição apresentando mais ferramentas que ilustrem sua performance e aplicabilidade.

5 Uso de Modelos Estatísticos na Análise de Dados de Reservatórios de Petróleo

Propõe-se um modelo que solucione o problema abordado em estudos geológicos referente a poços de petróleo. A fim de analisar o conjunto de dados, faz-se uso de modelagem estatística tendo como ferramentas a regressão logística e a análise discriminante de forma a definir um modelo preditivo para identificar tipos de rochas (litologias) favoráveis à acumulação de petróleo. Os modelos de classificação em pauta permitem realizar a avaliação de formações que definem a capacidade produtiva e a valoração das reservas de óleo e gás de poços de petróleo.

A justificativa para essa abordagem surge da extrema necessidade das empresas petrolíferas obterem informações de tipos de rochas (aqui rotulada como “fácies”) mediante a perfuração de poços, sendo estas caracterizadas como “rochas reservatório” e “rochas não-reservatório”. A primeira categoria apresenta como propriedade a capacidade de acumular e produzir petróleo. Em contra partida, a segunda categoria (não-reservatório) tem propriedade oposta à primeira. Dessa maneira, baseado nos tipos de rochas, sendo estas encontradas em cada nível de profundidade de um poço, pretende-se viabilizar a obtenção de informações mais precisas sobre os tipos de rochas com o intuito de reduzir os custos até então existentes quando se deseja analisar os poços de petróleo.

O uso da análise discriminante tem como objetivo obter a diferenciação das fácies considerando a existência de dois grupos que representam estas, para então propor um modelo que melhor discrimine os grupos. No caso da técnica de regressão logística, é possível classificar os tipos de litologias referentes as fácies reservatório e não-reservatório devido ao relacionamento existente entre a variável resposta binária e as variáveis explicativas que representam as curvas de perfis elétricos, os diferentes poços e o zoneamento destes. Como variáveis importantes desses modelos, capazes de classificar as fácies, são usados perfis geofísicos que são comuns aos poços e aos demais poços do campo onde

se deseja estimar os tipos litológicos.

Avalia-se assim a adequação dos diferentes tipos de modelos a fim de propor um modelo final. A utilização das técnicas de diagnóstico permite a identificação das observações que sejam influentes nas estimativas dos parâmetros dos modelos. Em particular, utilizar-se-á as técnicas de diagnóstico em nosso modelo de regressão logística.

O banco de dados tem um total de 1615 amostras de perfis geofísicos referentes às informações litológicas de três poços de petróleo. A partir do processamento e interpretação dos perfis geofísicos, são obtidas informações importantes a respeito das rochas contidas nos poços, como: litologia, espessura, porosidade, prováveis fluidos existentes nos poros e saturações (THOMAS et al., 2001). O termo "fácies", e seus derivados, é informal e normalmente usado para definir categorias segundo um critério previamente estabelecido. No que se segue, estes são resultados do artigo de Brito e Amaral (2007) publicado pela Revista Brasileira de Biometria.

5.1 Modelo de Regressão Logística Para Respostas Binárias

Dado que $\pi(x)$ varia entre 0 e 1, uma simples representação linear $x^T\beta$ para $\pi(\cdot)$ sobre todo o intervalo de x é impossível. O fato de ser impossível decorre da ocorrência de um determinado evento ser uma função não-linear das variáveis explicativas. Por essa razão, realiza-se a linearização mediante o uso da transformação logística $g(\pi)$, conhecida por $logit$, cujo parâmetro canônico é dado por

$$\eta = g(\pi) = \log[\pi/(1-\pi)], \tag{5.1}$$

em que a razão entre π e $1-\pi$ é denominada razão de chances.

Segundo Vittinghoff et al. (2005, p. 162) um dos mais significantes benefícios do modelo logístico é que os coeficientes de regressão são interpretados como o logaritmo da razão de chances.

O modelo geral de regresão logística é especificado como

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_1 + \beta_2 x_2 + \cdots + \beta_k x_k, \tag{5.2}$$

em que $x = (1, x_2, \ldots, x_k)^T$ contém os valores observados das $(p-1)$ variáveis explicativas.

Pretende-se analisar o relacionamento entre uma variável resposta binária, cuja representação apreende-se em indicar a ocorrência ou não de fácies reservatório, e a utilização de um conjunto de variáveis explicativas as quais representam curvas de perfis elétricos, diferentes poços e zoneamento destes. Por fim, deseja-se propor um modelo que prediza a variável observada de forma satisfatória.

5.2 Análise Discriminante

Considere M populações ou grupos $\pi_1, \ldots, \pi_M$, $M \geq 2$. Suponha que cada populacao π_i tem densidade de probabilidade $f_i(x)$ em $\mathbb{R}^p$. O objetivo da análise discriminante é alocar um indivíduo a um destes grupos com base em suas medidas x.

Fisher (1936), ao analisar o problema de discriminação entre as M populações, teve como principal objetivo encontrar uma função linear $b^T x$ de forma a maximizar a razão entre a soma dos quadrados entre os grupos e a soma dos quadrados dentro dos grupos. Ele não assume normalidade das observações populacionais, entretanto, assume implicitamente que a matriz de covariância das populações sejam iguais.

No caso em que a análise discriminante é aplicada a duas populações, tem-se que a função de classificação se resume a

$$w = (\bar{x}_1 - \bar{x}_2)^T \mathbf{S}^{-1} [x - \frac{1}{2}(\bar{x}_1 + \bar{x}_2)], \tag{5.3}$$

cuja denominação freqüente é dada pela função de classificação de Anderson (JOHNSON; WICHERN, 1992) devido ao fato de existir equivalência entre a regra de classificação de Fisher e a regra da mínima estimativa do custo de erro com probabilidades a $priori$ iguais e custos de erro de classificação iguais.

A aplicação da regra de classificação para dois grupos utilizando a função discriminante linear de Fisher, que foi definida em (5.3), indica que

$$w > 0 \Rightarrow x \in \pi_1,$$
$$w < 0 \Rightarrow x \in \pi_2.$$

Ao estudar o relacionamento da variável fácies com as demais variáveis, tal que essa variável seja representante de duas classes distintas, considera-se o grupo π_1 como sendo o representante da variável fácies não-reservatório e o grupo π_2 o representante da fácies reservatório.

Para a aplicação da discriminante linear logística, Anderson (1982) diz que o modelo discriminate logístico é uma descrição exata numa variedade de situações que incluem: densidades de classes condicionais normais multivariadas com matrizes de covariância iguais; distribuições discretas multivariadas seguindo um modelo log-linear com iguais termos de iteração entre os grupos; e a combinação de situações anteriores, ou seja, ambas variáveis contínuas e categóricas descrevendo cada amostra.

5.3 Análise de Dados em Reservatório de Petróleo

5.3.1 Introdução

Uma variável resposta e sete variáveis explicativas serão utilizadas na modelagem estatística e serão explicadas detalhadamente nos parágrafos seguintes.

A variável resposta *fácies* se caracteriza pela subdivisão em fácies reservatório e não-reservatório, onde a fácies reservatório indica a presença de petróleo.

A variável *Poço* representa os três poços analizados, sendo estes nomeados como *Poço*$_1$, *Poço*$_2$ e *Poço*$_3$. A variável *Zona* segue uma ordem que depende da profundidade e esta ainda define as possíveis litologias a ocorrerem na mesma. Mais especificamente, tem-se que para cada poço existe a presença dos diferentes tipos de zona, e ainda, em cada zona há os característicos tipos de litologia. De modo geral, existem 11 tipos de zonas que descrevem a estrutura do poço, mas estas foram reagrupadas e, por consegüinte, os fatores a serem utilizados para a análise estatística são $Zona_{1A}$, $Zona_{1B}$, $Zona_2$, $Zona_{3A}$, $Zona_{3B}$ e $Zona_4$.

A variável raios gama (*GR*) mede primariamente a radioatividade natural das rochas, ou seja, a radioatividade total da formação geológica. É utilizada para a identificação da litologia, a identificação de minerais radioativos e para o cálculo do volume de argilas ou argilosidade. Como regra geral, quanto mais radioativa a rocha menor a sua granulometria.

A variável indução (*ILD*) fornece a leitura aproximada da resistividade da rocha por meio da medição de campos elétricos e magnéticos induzidos. De forma geral, rochas porosas com óleo têm resistividade alta e com água salgada a resistividade é baixa.

A variável densidade (*RHOB*) é uma medida de densidade eletrônica que detecta os raios gama defletidos pelos elétrons dos elementos das rochas, após terem sido emitidos por uma fonte colimada situada dentro do poço. Além disto, através da densidade é possível o cálculo da porosidade e a identificação das zonas de gás.

Uma outra variável definida como sônico (*DT*) mede a diferença nos tempos de trânsito de uma onda mecânica por meio das rochas. É utilizado para estimativas de porosidade, correlação poço a poço, estimativas do grau de compactação das rochas ou estimativas das constantes elásticas, detecção de fraturas e apoio às sísmicas para a elaboração do sismograma sintético.

A variável porosidade neutrônica (*NPHI*) é uma medida de densidade da rocha. É medida sob o aspecto da emissão de nêutrons. Os perfis mais antigos medem a quantidade de raios gama de captura após excitação artificial mediante bombardeio dirigido de nêutrons rápidos. Os mais modernos medem a quantidade de nêutrons epitermais ou termais da rocha após bombardeio. São utilizados para estimativa de porosidade, litologia e detecção de hidrocarbonetos leves ou gás.

5.3.2 Análise descritiva das variáveis

Antes de ajustar os modelos, foi realizada uma análise da estrutura de correlação entre as variáveis. Na Tabela 5.1 tem-se a matriz de correlação entre as variáveis do modelo. A matriz indica a presença de uma forte correlação de algumas das variáveis explicativas entre si.

Na Tabela 5.2 temos a análise descritiva das variáveis analisadas em cada poço e a análise conjunta destes. Ao particionar os poços a fim de analisá-los individualmente, observa-se que estes apresentam as mesmas características entre as variáveis.

Depois de alguns modelos iniciais, verificou-se a necessidade de aplicar uma transformação na variável *ILD*. A transformação segue por meio da aplicação do logaritmo devido ao fato de originalmente essa variável não realizar um bom ajuste nas caudas do modelo.

Em se tratando de geoestatística para análise espacial dos dados, os geólogos aplicam uma transformada para a variável ILD. A transformada aplicada utiliza-se da distribuição log-normal tri-paramétrica pelo fato de existir casos em que a distribuição log-normal bi-paramétrica não é simétrica e, por conseguinte, não é log-normal. Maiores detalhes sobre a distribuição log-normal tri-paramétrica podem ser obtidos mediante Hill (1963), Hirose (1997), Wingo (1984), entre outros.

Tabela 5.1: Matriz de correlação entre as variáveis propostas para definição do modelo.

Covariáveis	GR	DT	log(ILD)	RHOB	NPHI
GR	$1,00$				
DT	$0,81$	$1,00$			
log(ILD)	$-0,74$	$-0,61$	$1,00$		
RHOB	$-0,39$	$-0,76$	$0,20$	$1,00$	
NPHI	$0,84$	$0,94$	$-0,62$	$-0,71$	$1,00$

5.3.3 Modelo de Regressão Logística

Na Tabela 5.3 é apresentada a descrição da análise de desvio por meio do processo de definição de modelos significativos obtidos a cada adição, sendo o modelo final apresentado por

$$\begin{aligned}\Im \;=\; & \beta_0+\beta_1\times GR+\beta_2\times\log(ILD)+\beta_3\times RHOB+\\ & +\beta_4\times \textit{Poço}+\beta_5\times \textit{Zona}+\beta_6\times \textit{Poço}\times \textit{Zona}.\end{aligned} \tag{5.4}$$

As estimativas dos parâmetros deste modelo estão na Tabela 5.4.

Ao verificar as estimativas da razão de chances para cada efeito das variáveis na Tabela 5.4, tem-se que os valores estimados das razões de chances $\hat{\Psi}(cte)$, $\hat{\Psi}(\log(ILD))$, da associação conjunta de $\hat{\Psi}(Zona_{1A},\ Zona_{1B})$, $\hat{\Psi}(Zona_{1A},\ Zona_2)$, $\hat{\Psi}(Zona_{1A},\ Zona_{3A})$, $\hat{\Psi}(Zona_{1A},\ Zona_{3B})$, $\hat{\Psi}(Zona_{1A},\ Zona_4)$ e da associação conjunta das interações do modelo referentes a $\hat{\Psi}(\textit{Poço}_2\times Zona_2)$, $\hat{\Psi}(\textit{Poço}_2\times Zona_{3A})$, $\hat{\Psi}(\textit{Poço}_2\times Zona_{3B})$, $\hat{\Psi}(\textit{Poço}_2\times Zona_4)$ apresentam uma forte contribuição na relação desses efeitos com a variável resposta, aqui referida como fácies reservatório. Ou seja, indicam de forma particular a chance de se definir fácies reservatório quando analisada uma variável, em particular, e considerada as demais variáveis como constantes. É observado ainda uma fraca, mas ainda positiva, associação dos efeitos das variáveis por meio das seguintes razões de chances: $\hat{\Psi}(\textit{Poço}_1,\textit{Poço}_3)$, $\hat{\Psi}(\textit{Zona}_{1A},\textit{Zona}_{3B})$ e $\hat{\Psi}(\textit{Poço}_3\times Zona_{3B})$. Quanto a estimativa da razão de chances para os efeitos das variáveis referentes às $\hat{\Psi}(RHOB)$, $\hat{\Psi}(\textit{Poço}_3\times Zona_{1B})$, $\hat{\Psi}(\textit{Poço}_3\times Zona_2)$, $\hat{\Psi}(\textit{Poço}_3\times Zona_{3A})$ e $\hat{\Psi}(\textit{Poço}_3\times Zona_4)$, tem-se que estas apresentam uma contribuição na variável resposta correspondente a um fator protetor, ou seja, uma associação negativa de contribuição da variável.

Tabela 5.2: Análise descritiva dos poços de maneira geral e individual.

	Variável	Mínimo	1º Quartil	Mediana	Média	3º Quartil	Máximo	D. Padrão	Assimetria	Curtose
Geral	GR	30,300	47,900	64,400	70,560	89,950	141,800	25,874	0,574	2,257
	DT	59,500	74,550	84,800	85,430	95,500	118,600	12,669	0,203	2,063
	log(ILD)	0,000	1,131	1,758	1,871	2,510	5,011	0,957	0,442	2,834
	RHOB	2,130	2,380	2,430	2,428	2,480	2,630	0,078	−0,417	3,107
	NPHI	8,100	19,700	25,300	25,240	30,700	39,500	6,507	−0,033	2,091
Poço 1	GR	34,600	52,600	70,000	72,880	88,900	135,800	23,909	0,477	2,361
	DT	59,500	79,600	91,100	89,870	100,000	118,600	12,619	−0,172	2,212
	log(ILD)	0,000	1,030	1,629	1,846	2,532	5,011	1,102	0,720	3,201
	RHOB	2,130	2,333	2,390	2,392	2,450	2,600	0,082	−0,119	2,840
	NPHI	8,100	22,000	27,100	26,340	31,200	39,500	6,191	−0,409	2,601
Poço 2	GR	31,400	45,500	62,600	70,450	94,150	141,800	27,693	0,545	2,051
	DT	59,700	72,250	79,300	81,090	90,100	104,600	10,552	0,203	1,868
	log(ILD)	0,531	1,224	1,841	1,859	2,402	3,721	0,761	0,341	2,312
	RHOB	2,310	2,420	2,470	2,459	2,500	2,630	0,060	−0,233	2,685
	NPHI	9,900	19,200	23,700	24,330	29,400	38,800	6,206	0,135	2,050
Poço 3	GR	30,300	47,900	60,700	68,690	87,920	140,200	25,111	0,716	2,498
	DT	61,200	75,030	85,550	86,840	98,400	115,200	13,405	0,191	1,864
	log(ILD)	0,000	1,099	1,758	1,906	2,771	3,910	1,033	0,151	2,061
	RHOB	2,150	2,370	2,420	2,422	2,470	2,610	0,079	−0,291	3,126
	NPHI	10,200	19,520	25,250	25,410	31,380	39,200	6,965	0,037	1,938

Finalmente, tem-se que para as razões de chances $\hat{\Psi}(GR)$, $\hat{\Psi}(Poço_1, Poço_2)$, $\hat{\Psi}(Poço_2 \times Zona_{1B})$, a associação com a variável resposta quase não existe, pois é importante lembrar que, quando a razão de chances é um, pode-se considerar que a associação com a variável resposta seja nula.

Estando definido o modelo e sendo ele, também, o que apresenta menor $AIC = 1130$, o necessário agora é analisar os resíduos por meio dos métodos de diagnóstico para verificar a adequacidade do modelo proposto.

Segundo Hosmer e Lemeshow (1989, p. 157), uma conseqüência prática ao avaliar os pontos de alavanca na regressão logística é que para interpretar corretamente um valor particular de alavanca, precisa-se saber se o valor estimado de probabilidade é menor que $0,1$ ou maior que $0,9$. Se a probabilidade estimada estiver entre $0,1$ e $0,9$ então a alavanca dará um valor que pode ser referido como distância. Assim, utilizaria-se da mesma consideração da regressão linear em que a alavanca é uma função monótona de incremento da distância da matriz de covariância padronizada para a média. Quando um estimador de probabilidade não está nos limites do intervalo $(0,1,\ 0,9)$, então o valor da alavanca não pode ser considerado medida de distância no sentido que isto implica altos valores.

As Figuras de 5.1(A) a 5.1(C) apresentam alguns gráficos de diagnóstico considerados

Tabela 5.3: Modelo obtido após realização da análise de desvio para definição das variáveis presentes no mesmo.

Modelo	Desvio	Diferença	g.l.	p-valor	Testando
Constante (cte)	$1922,37$	-	-	-	-
+ GR	$1239,42$	$682,95$	1	$1,526 \times 10^{-150}$	GR
+ log(ILD)	$1178,63$	$60,82$	1	$6,332 \times 10^{-15}$	$\log$(ILD)| GR
+ Zona	$1110,52$	$68,11$	5	$2,531 \times 10^{-13}$	Zona| GR $+\log$(ILD)
+ Poço $\times$ Zona	$1080,25$	$30,27$	12	$0,002545$	Poço$\times$Zona| GR $+\log$(ILD) +Poço + Zona
+ RHOB	$1071,39$	$8,86$	1	$0,002912$	Poço $\times$ Zona | GR $+\log$(ILD) + Poço +Zona + RHOB

Tabela 5.4: Estimativas dos parâmetros e respectivos desvio padrão e razão de chances referentes ao modelo logístico com efeitos principais para explicar a ocorrência de fácies reservatório.

Efeito	Estimativa	D. Padrão	Razão de Chances $(\hat{\Psi})$
Constante	$11,443539(3,296)$	$3,472115$	$9,329663 \times 10^4$
GR	$-0,050913(-7,903)$	$0,006442$	$0,950361$
$\log$ *(ILD)*	$1,102887(5,844)$	$0,188718$	$3,012853$
RHOB	$-3,832972(-2,978)$	$1,287028$	$2,164518 \times 10^{-2}$
$Poço_2$	$-0,016636(-0,042)$	$0,400497$	$0,983501$
$Poço_3$	$0,500014(1,356)$	$0,368781$	$1,648745$
$Zona_{1B}$	$0,776408(1,780)$	$0,436123$	$2,173650$
$Zona_2$	$1,009750(2,359)$	$0,428093$	$2,744915$
$Zona_{3A}$	$1,179471(2,262)$	$0,521477$	$3,252653$
$Zona_{3B}$	$0,446934(1,023)$	$0,436856$	$1,563511$
$Zona_4$	$1,379191(2,843)$	$0,485044$	$3,971686$
$Poço_2 \times Zona_{1B}$	$0,070685(0,120)$	$0,591231$	$1,073243$
$Poço_3 \times Zona_{1B}$	$-0,864786(-1,460)$	$0,592236$	$0,421142$
$Poço_2 \times Zona_2$	$1,254208(2,126)$	$0,589878$	$3,505062$
$Poço_3 \times Zona_2$	$-0,810829(-1,341)$	$0,604504$	$0,444897$
$Poço_2 \times Zona_{3A}$	$0,898854(1,243)$	$0,723076$	$2,456778$
$Poço_3 \times Zona_{3A}$	$-0,286676(-0,407)$	$0,704204$	$0,750755$
$Poço_2 \times Zona_{3B}$	$1,285961(2,040)$	$0,630259$	$3,618142$
$Poço_3 \times Zona_{3B}$	$0,490592(0,796)$	$0,637715$	$1,633282$
$Poço_2 \times Zona_4$	$1,202030(1,866)$	$0,644263$	$3,326865$
$Poço_3 \times Zona_4$	$-1,034729(-1,679)$	$0,616433$	$0,355323$

por Hosmer e Lemeshow (1989, pp. 160-161) como gráficos base para a análise de diagnóstico do modelo de regressão logística. Na Figura 5.1(A) tem-se o gráfico referente aos resíduos do componente desvio padronizado com relação aos valores ajustados de forma a apresentar também a distribuição dos resíduos no intervalo $[-2;2]$. Ao verificar o gráfico da Figura 5.1(B), referente aos pontos de alavanca versus valores ajustados, observa-se a aparente existência de dispersão de alguns pontos.

O gráfico da distância de Cook, referente à Figura 5.1(C), tem como função descrever a influência de pontos no modelo. Assim, para a Figura 5.1(C), não há indícios de pontos influentes por causa dos baixos valores encontrados ao calcular a distância de Cook.

Apesar das informações descritas pelos gráficos, ao verificar o desvio do modelo tem-se que este apresenta valor baixo ao relacioná-lo com os graus de liberdade. Ou seja, há indícios de subparametrização devido ao fato do desvio ser $D(y;\hat{\mu}) = 1100$ e os graus de liberdade do modelo corresponderem a 1614. O gráfico normal de probabilidades com

envelopes para o resíduo de componente do desvio (Figura 5.1(D)) não apresenta indícios de problemas sérios da suposição de distribuição binomial para a variável resposta, pois a maioria dos pontos apresentam-se dentro do envelope.

Na Tabela 5.5 são apresentadas as medidas de diagnóstico: resíduo de Pearson padronizado ($R^*_{P_i}$), resíduo do componente desvio padronizado ($R^*_{D_i}$), distância de Cook (LD_i) e alavancagem ($\hat{h_{ii}}$), relacionando estes com as probabilidades ajustadas. As situações consideradas na Tabela 5.5 não correspondem a situação apresentada no gráfico quando analisamos o intervalo de $\hat{\pi}_i$. Hosmer e Lemeshow (1989, p. 161) consideram, para a análise do ponto de alavanca, o uso do valor crítico da distribuição qui-quadrado com um grau de liberdade ao nível de significância de 95%. Sendo assim, devido aos valores dos pontos de alavanca serem menores que o valor crítico de $3,84$, então, não há como considerar

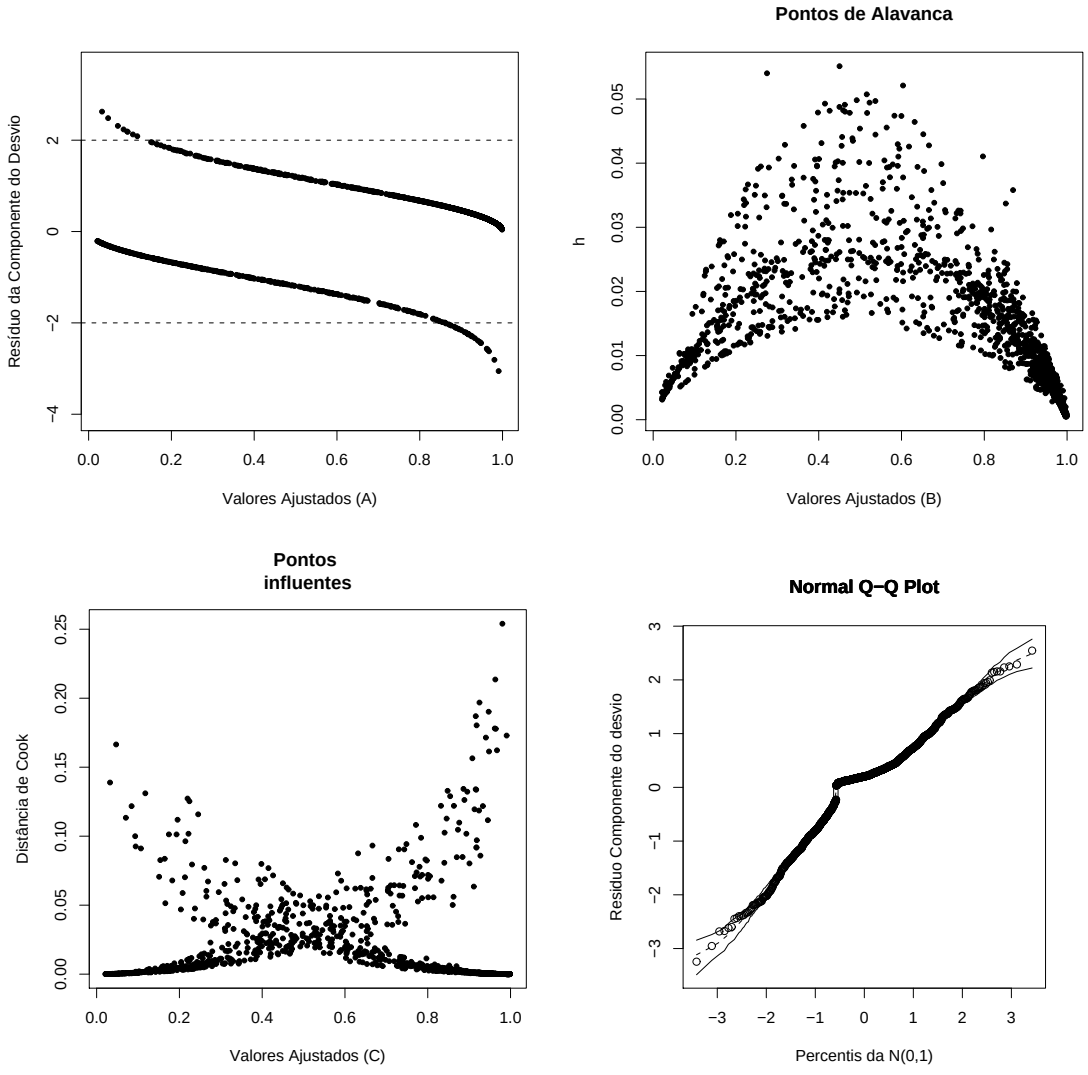

Figura 5.1: Análise dos resíduos do modelo ajustado.

Tabela 5.5: Possíveis valores para as medidas de diagnóstico $R^*_{P_i}$, $R^*_{D_i}$, LD_i e h_i com cinco regiões definidas segundo as probabilidades ajustadas.

Diagnóstico	Probabilidade Ajustada				
	$0,0$ - $0,1$	$0,1$ - $0,3$	$0,3$ - $0,7$	$0,7$ - $0,9$	$0,9$ - $1,0$
$R^*_{P_i}$ ou $R^*_{D_i}$	Menor/Maior	Moderado	Moderado a menor	Moderado	Menor/Maior
LD_i	Menor	Maior	Moderado	Maior	Menor
$\hat{h}_{ii}$	Menor	Maior	Moderado a menor	Maior	Menor

presença de valores que afetem o ajuste do modelo.

Ao ser utilizado um ponto de corte em 50% para o modelo logístico, observa-se uma taxa de erro em torno de $15,47\%$. Na Tabela 5.6 consta a matriz de confusão para o modelo logístico.

Tabela 5.6: Matriz de confusão para o modelo de regressão logística.

População Verdadeira	Classificação Realizada	
	0	1
0	318	112
1	138	1047

Como passo seguinte, avalia-se o poder de classificação do modelo, com todas as observações presentes, mediante o método simples de classificação logístico e, também, por meio do método de curva ROC.

Segundo Louzada-Neto e Martinez (2000), quando o teste sob investigação produz uma resposta sob a forma de uma variável categórica ordinal ou contínua, emprega-se uma regra de decisão baseada em buscar um ponto de corte que resume tal quantidade em uma resposta dicotômica. Desta forma, para diferentes pontos de corte dentro da amplitude dos possíveis valores que o teste sob investigação produz, pode-se estimar sensibilidades e especificidades. Sabe-se, ainda, que a curva ROC tem como vantagem a avaliação de métodos de diagnóstico por meio do seu gráfico. Baseado nisto, deseja-se avaliar o desempenho do modelo logístico, considerado aqui como um teste de diagnóstico, de acordo com o conjunto de suas possíveis respostas por meio da sua representação visual direta. Na Figura 5.2 está descrita a curva ROC para o modelo logístico proposto, onde se observa que a curva caracteriza uma boa descrição do poder de classificação do modelo. Mediante a regra dos trapézios, a área sob a curva ROC é estimada em $0,91$ indicando

que o modelo tem boa capacidade preditiva. A curva ROC não utiliza conjunto de teste e treinamento em suas análises de modelos.

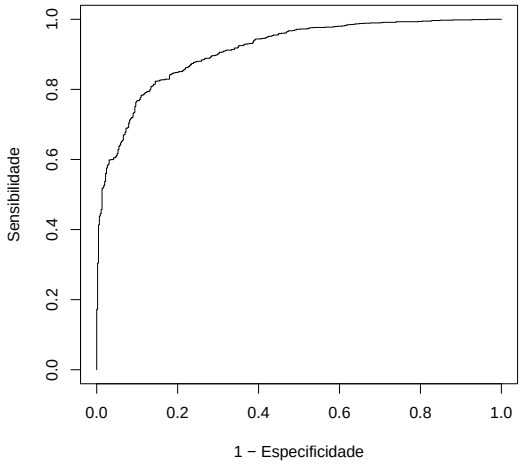

Figura 5.2: Curva ROC do modelo logístico proposto.

Por meio de uma amostra de teste é possível estimar de forma honesta a taxa de erro. A amostra de teste consiste em dividir o conjunto de dados em duas amostras independentes. Usa-se uma dessas amostras para obter a função de classificação e a amostra seguinte servirá como teste para estimação da taxa de erro. Sugere-se usar $\frac{2}{3}$ dos dados para treinamento e $\frac{1}{3}$ para teste. Assim, verifica-se a taxa de má-classificação por

Tabela 5.7: Matriz de confusão para o modelo de regressão logística usando o conjunto de teste.

População Verdadeira	Classificação Realizada	
	0	1
0	116	53
1	37	332

meio do uso de um conjunto de teste e um conjunto de treinamento para a formação do modelo, onde $\frac{1}{3}$ corresponde ao conjunto de teste. As amostras referentes ao conjunto de treinamento apresentam 303 observações correspondendo ao grupo não reservatório e 774 observações para o grupo reservatório. A Tabela 5.7 descreve as classificações obtidas por meio do conjunto com 538 amostras de teste para o modelo proposto. A partir desse conjunto de teste foi obtida uma taxa de má-classificação em torno de $16,73\%$ para o modelo logístico.

5.3.4 Modelo de Análise Discriminante

Utilizando o mesmo conjunto de teste obtido para a análise do modelo logístico, com probabilidade a priori para o grupos de forma proporcional, ou seja, grupo 1 com probabilidade aproximada $0,7176$ e grupo zero com probabilidade aproximada $0,2823$. A matriz de confusão referente ao modelo proposto (5.4) apresenta uma taxa de erro de $14,87\%$, ver Tabela 5.8.

Tabela 5.8: Matriz de confusão para o modelo de análise discriminante baseado no modelo (5.4).

População Verdadeira	Classificação Realizada	
	0	1
0	114	39
1	41	344

Tabela 5.9: Valores médios das variáveis por grupo no modelo discriminante linear e a estimativa dos coeficientes.

Variável	Média dos grupos		Coeficiente
	Não Reservatório	Reservatório	
GR	$96,234320$	$60,726490$	$-0,040467$
$\log(ILD)$	$1,056693$	$2,183419$	$0,349664$
$RHOB$	$2,411254$	$2,433346$	$-3,734885$
$Poço_2$	$0,339934$	$0,396641$	$-0,363749$
$Poço_3$	$0,349835$	$0,329457$	$0,170243$
$Zona_{1B}$	$0,290429$	$0,080103$	$-0,048403$
$Zona_2$	$0,161716$	$0,229974$	$0,232578$
$Zona_{3A}$	$0,085809$	$0,157623$	$0,070476$
$Zona_{3B}$	$0,125412$	$0,143411$	$-0,285982$
$Zona_4$	$0,108911$	$0,236434$	$0,173034$
$Poço_2 \times Zona_{1B}$	$0,105611$	$0,025840$	$0,406046$
$Poço_3 \times Zona_{1B}$	$0,105611$	$0,024548$	$-0,523776$
$Poço_2 \times Zona_2$	$0,049505$	$0,073643$	$0,977192$
$Poço_3 \times Zona_2$	$0,052805$	$0,085271$	$-0,407053$
$Poço_2 \times Zona_{3A}$	$0,019802$	$0,062015$	$1,030522$
$Poço_3 \times Zona_{3A}$	$0,029703$	$0,046512$	$0,211381$
$Poço_2 \times Zona_{3B}$	$0,033003$	$0,041344$	$1,255827$
$Poço_3 \times Zona_{3B}$	$0,039604$	$0,058139$	$0,191865$
$Poço_2 \times Zona_4$	$0,026403$	$0,152455$	$0,934092$
$Poço_3 \times Zona_4$	$0,049505$	$0,055555$	$-0,318819$

A média das variáveis por grupo para o modelo proposto pela equação (5.4) é dada pela Tabela 5.9. Ao serem verificados os coeficientes estimados para o modelo representado pela equação 5.4, observa-se que os coeficientes das variáveis $log(ILD)$, $Poço_3$,

Zona$_2$, *Zona*$_4$, *Poço*$_2\times$ *Zona*$_{1B}$, *Poço*$_3\times$ *Zona*$_{1B}$, *Poço*$_2\times$ *Zona*$_2$, *Poço*$_2\times$ *Zona*$_{3A}$, *Poço*$_3\times$ *Zona*$_{3A}$, *Poço*$_2\times$ *Zona*$_{3B}$, *Poço*$_3\times$ *Zona*$_{3B}$ e *Poço*$_2\times$ *Zona*$_4$ apresentam maior ênfase em suas características, ou seja, auxiliam de forma positiva na classificação dos grupos no modelo de análise discriminante.

5.3.5 Análise Final do Modelo Proposto

Após tomar conhecimento das características das variáveis utilizadas para definição do modelo, por meio da análise do desvio, foi definido como o melhor modelo o que continha as variáveis *GR*, $\log(ILD)$, *RHOB*, *Poço* e *Zona*. O modelo proposto (5.4) foi abordado também na análise discriminante com o intuito de comparar as taxas de má-classificação entre os dois modelos utilizados. A utilização da curva ROC objetivou avaliar o modelo de regressão logística. Dessa forma, observa-se que ao utilizar a curva ROC, o modelo logístico apresentou boa capacidade preditiva pois a área sob a curva foi estimada em $0,91$ pela regra dos trapézios. O uso da curva ROC vem propor então a utilização do ponto de corte definido por esta no modelo logístico ao invés de utilizar o corte de 50% da regessão logística para a classificação da variável fácies.

Para verificar a eficiência entre os métodos, observou-se qual apresentava menor taxa de má-classificação das observações. Sendo assim, após serem obtidas as matrizes de confusão para regressão logística (Tabela 5.7), cuja taxa de má-classificação corresponde a $16,73\%$, e análise discriminante (Tabela 5.8), cuja taxa de má-classificação é de $14,87\%$, observou-se que a análise discriminante apresenta menor taxa de má-classificação.

Segundo Cox e Snell (1989), seria possível, então, decidir que o melhor método para aplicação do conjunto de dados seria usar o modelo logístico devido a não necessidade de obtermos sub-populações. Mas, dado os resultados obtidos, pode-se concluir que os dois métodos se mostram eficientes quando modelado o conjunto de dados.

5.4 Considerações Finais

O presente Capítulo seguiu uma proposta que vise analisar o conjunto de dados referente à reservatórios de petróleo mediante método probabilístico. Mediante isto, sugeriu-se dois tipos de modelos, sendo o primeiro referente ao modelo de regressão logística e o segundo, referente ao modelo de análise discriminante. Existem diferentes métodos para análise dos dados aqui abordados, a exemplo de redes neurais, porém, estes modelos seriam uma proposta inicial com o enfoque de probabilidade. A partir do que foi apre-

sentado neste capítulo, recomenda-se um estudo mais aprofundado com a finalidade de melhor interpretar as variáveis disponíveis para a modelagem e, assim, obter modelos mais apropriados.

Diante das considerações de Cox e Snell (1989) para interpretação de modelos, poderia ser feito o uso de ambos modelos para análise das litologias correspondentes aos reservatórios de petróleo através da equação (5.4). E, uma sugestão adicional seria o uso do ponto de corte obtido através da curva ROC com o intuito de melhorar o critério de classificação quando usado o modelo de regressão logística.

APÊNDICE A – Algoritmo para Avaliação Numérica dos Momentos da Beta Normal

```
% Momentos da beta normal

t1 = cputime;

a=0.5;
b=1.5;
p=0;
q=1;

f= @(x) normpdf(x,p,q);

format short eng

disp('r              Momentos')

for r=1:6
    % Implements Simpson's rule using for loop.
    m=10000;

    lim_inf=-10^3;
    lim_sup= 10^3;
    if (m/2)~=floor(m/2)
        disp('m must be even'); break
    end
    h=(lim_sup-lim_inf)/m;
    x=lim_inf;

    % avalia y1
     G=normcdf(x,p,q);
     fbt=(f(x)/beta(a,b))*(G^(a-1))*(1-G)^(b-1);
```

```
     y1= (x^r)*fbt;

    s=0;
    for j=2:2:m

        % avalia ym
        x=lim_inf+(j-1)*h;

        G=normcdf(x,p,q);
        fbt=(f(x)/beta(a,b))*(G^(a-1))*(1-G)^(b-1);
        ym= (x^r)*fbt;

        % avalia yh
        x=lim_inf+j*h;

        G=normcdf(x,p,q);
        fbt=(f(x)/beta(a,b))*(G^(a-1))*(1-G)^(b-1);
        yh=(x^r)*fbt;

        % soma
        s=s+y1+4*ym+yh;
        y1=yh;
    end
    mu(r)=s*h/3;;

disp([num2str(r) '           '  num2str(mu(r))]);
end

var = mu(2)-(mu(1))^2;
ass = (mu(3)-3*mu(1)*mu(2)+2*mu(1)^3)/var^(3/2);
kur = (mu(4)-4*mu(1)*mu(3)+6*(mu(1)^2)*mu(2)-3*mu(1)^4)/var^2;

Delta1=(cputime-t1)/60;

disp('                                                          ')
disp('   ----------------------------------------------------')
disp(['  var:        ' num2str(var)])
disp(['  ass:        ' num2str(ass)])
disp(['  kur:        ' num2str(kur)])
disp('   ----------------------------------------------------')
disp(['  t: ' num2str(Delta1) 'min'])
```

Referências Bibliográficas

ANDERSON, J. A. **Handbook of Statistics**. P. R. Krishnaiah and L. N. Kanal: North-Holland Publishing Company, v. 2, cap. Logistic Discrimination, p. 169-191, 1982.

APPELL, P.; FÉRIET, J. Kampé de. **Fonctions hypergéometriques et hypersphériques: polynomes d'Hermite**. [S.l.]: Paris, 1926.

BAHADUR, R. R.; RAO, R. R. On deviations of the sample mean. **Annals of Mathematics and Statistics**, v. 31, p. 1015–1027, 1960.

BALAKRISHNAN, N.; NEVZOROV, V. B. **A Primer on Statistical Distributions**. [S.l.]: John Wiley and Sons, Inc., New Jersey, cap. 14, p. 127–132, 2003.

BARAKAT, H. M.; ABDELKANDER, Y. H. Computing the moments of order statistics from noindentical random variables. **Statistical Methods and Applications**, v. 13, n. 1, p. 15–26, 2004.

BARNDORFF-NIELSEN, O.; COX, D. R. Edgeworth and saddlepoint approximations with statistical applications. **Journal Royal Statistical. Society Serie B**, v. 41, p. 279–312, 1979.

BARRETO-SOUZA, W.; SANTOS, A. H. S.; CORDEIRO, G. M. The beta generalized exponential distribution. **a aparecer no JSCS**, 2009.

BARTLETT, M. S. Properties of sufficiency and statistical tests. **Proceedings of the Royal Society of London A**, v. 160, p. 268–282, 1937.

BLACKWELL, D.; HODGES, J. L. The probability in the extreme tail of a convolution. **Annals of Mathematics and Statistics**, v. 31, p. 1113–1120, 1959.

BRITO, R. S.; AMARAL, G. J. A. Uso de modelos estatísticos na análise de dados de reservatórios de petróleo. **Revista Brasileira de Biometria, São Paulo**, v. 25, n. 3, p. 93–107, 2007.

BRITO, R. S.; CORDEIRO, G. M. Comparação das expansões de Edgeworth, Lugannani-Rice, Daniels e Cordeiro-Ferrari com aplicações estatísticas. **submetido a Revista Brasileira de Biometria**, p. 1–29, 2009.

BURY, K. **Statistical Distributions in Engineering**. [S.l.]: New York, 1999.

CORDEIRO, G. M. **Introdução à Teoria Assintótica - Livro Texto do 22º Colóquio Brasileiro de Matemática**. [S.l.]: Rio de Janeiro, 1999.

CORDEIRO, G. M.; CRISTINO, C. The beta generalized Rayleigh distribution. **em preparação**, 2009.

CORDEIRO, G. M.; FERRARI, S. L. P. A modified score test statistics having chi-squared distribution to order n^{-1}. **Biometrika**, v. 78, p. 573–582, 1991.

CORDEIRO, G. M.; FERRARI, S. L. P. Generalized Bartlett correction. **Commun. Statist.-Theory Meth.**, v. 27, p. 509–527, 1998.

CORDEIRO, G. M.; FERRARI, S. L. P.; CYSNEIROS, A. H. M. A. A formula to improved score test statistics. **Journal of Statistical Computation and Simulation**, v. 62, p. 123–136, 1998.

CORDEIRO, G. M.; NADARAJAH, S. Closed form expressions for moments of a class of beta generalized distributions. **pre-print**, p. 1–14, 2009.

COX, D. R.; SNELL, E. J. **Analysis of Binary Data**. [S.l.]: UK, 1989.

CRAMÉR, H. Sur un nouveu théorèm-limite de la théorie des probabilités. **Actualites Scientifiques et Industrielles, Paris: Hermann et Cie**, 1938.

CYSNEIROS, A. H. M. A.; CORDEIRO, G. M. On improving the χ^2 approximation of score tests in location-scale nonlinear models. **Communications in Statistics - Theory and Methods**, v. 31, n. 10, p. 1709–1732, 2002.

CYSNEIROS, F. J. A.; CORDEIRO, G. M.; CYSNEIROS, A. H. M. A. Corrected maximum likelihood estimators im heteroscedastic symmetric nonlinear models. **a aparecer no Journal of Statistical Computation and Simulation**, 2009.

DANIELS, H. E. Saddlepoint approximations in statistics. **The Ann. Math. Statistics**, v. 25, p. 631–650, 1954.

DANIELS, H. E. Exact saddlepoint approximations. **Biometrika**, v. 67, p. 59–63, 1980.

DANIELS, H. E. Tail probability approximations. **International Statistical Review**, v. 55, p. 37–48, 1987.

DAVISON, A. C. Biometrika centenary: Theory and general methodology. **Biometrika**, v. 88, p. 13–52, 2001.

DENNIS, B.; PATIL, G. P. The gamma distribution and the weighted multimodal gamma distributions as models od population abundance. **Mathematical Biosciences**, v. 68, p. 187–212, 1984.

ERDÉLYI, A. Über einige bestimmte integrale, in denen die whittakerschen $m_{k,m}$–funktionen auftreten. **Mathematische Zeitschrift**, v. 40, p. 693–702, 1936.

ESSCHER, F. The probability function in the collective theory of risk. **Skand. Akt. Tidsskr. (Scand. Actuarial J.)**, v. 15, p. 175–195, 1932.

EUGENE, N.; LEE, C.; FAMOYE, F. Beta-normal distribution and its applications. **Commun. Statist.-Theory and Methods**, v. 31, p. 497–512, 2002.

EXTON, H. **Handbook of Hypergeometric Integrals: Theory, Applications, Tables, Computer Programs**. [S.l.]: New York, 1978.

FERRARI, S. L. P.; CRIBARI-NETO, F. Beta regression for modelling rates and proportions. **Journal of Applied Statistics**, v. 31, p. 799–815, 2004.

FINNER, H.; DICKHAUS, T.; ROTERS, M. Asymptotic tail properties of “Student’s” t-distribution. **Communications in Statistics - Theory and Methods**, v. 37, p. 175–179, 2008.

FISHER, R. A. Expansion of "Student’s"integral in powers of n^{-1}. **Metron**, v. 5, p. 109–120, 1925.

FISHER, R. A. The use of multiple measurements in taxonomic problems. **Annals of Eugenics**, v. 7, p. 179–188, 1936.

GOUTIS, C.; CASELLA, G. Explaining the saddlepoint approximation. **The American Satistician**, v. 53, p. 216–224, 1999.

GRADSHTEYN, I. S.; RYZHIK, I. M. **Table of integrals, series, and products**. [S.l.]: San Diego, 2000.

GRAHAM, V. A.; HOLLANDS, K. G. T. Method to generate synthetic hourly solar radiation globally. **Solar Energy**, v. 44, p. 333–341, 1990.

HILL, B. M. The three-parameter lognormal distribution and bayesian analysis of a point-source-epidemic. **Journal of the American Statistical Association**, v. 58, p. 72–84, 1963.

HINKLEY, D. V.; REID, N.; SNELL, E. J. **Statistical Theory and Modelling**. [S.l.]: In honour of Sir David Cox, 1990.

HIROSE, H. Maximun likelihood parameter estimation in the three-parameter log-normal distribution usinf the continuation method. **Computational Statistics and Data Analysis**, v. 24, p. 139–152, 1997.

HOSMER, J. D.; LEMESHOW, S. **Applied Logistic Regression**. [S.l.]: John Wiley and Sons, 1989.

JANARDAN, H. G.; PADMANABHAN, G. Double bounded beta distribution for hydrologic variables. In: **Proc. 17th Annual Pittsburg Conference (parte 3)**. [S.l.: s.n.], 1986. v. 17, p. 1107–1111.

JOHNSON, N. L.; KOTZ, S.; BALAKRISHNAN, N. **Continuous Univariate Distribution, v. 1**. [S.l.]: A Wiley-Interscience Publication, 1995a.

JOHNSON, N. L.; KOTZ, S.; BALAKRISHNAN, N. **Continuous Univariate Distribution, v. 2**. [S.l.]: A Wiley-Interscience Publication, 1995b.

JOHNSON, R. A.; WICHERN, D. W. **Applied Multivariate Statistical Analysis**. Third edition. [S.l.]: Englewood Cliffs, p. 524, 1992.

KAKIZAWA, Y. Higher order monotone Bartlett-type adjustment for some multivariate test statistics. **Biometrika**, v. 83, p. 923–927, 1996.

KENDALL, M. G. **The Advanced Theory of Statistics**. [S.l.]: London, v. 1, p. 158, 1945.

KOLASSA, J. E. **Series Approximation Methods in Statistics - Lecture Notes in Statistics 88**. Second. [S.l.: s.n.], 1997.

KONG, L.; LEE, C.; SEPANSKI, J. H. On the properties of beta-gamma distribution. **Journal of Modern Applied Statistical Methods**, v. 6, n. 1, p. 187–211, 2007.

KOTZ, S.; NADARAJAH, S. **Extreme Value Distribution. Theory and Applications**. [S.l.]: Imperial College Press, 2000.

LAURICELLA, G. Sulla funzioni ipergeometriche a più variabili. **Rend. Circ. Math. Palermo**, v. 7, p. 111–158, 1893.

LOUZADA-NETO, F.; MARTINEZ, E. Z. Metodologia estatística para testes diagnósticos e laboratoriais com respostas dicotomizadas. **Revista de Matemática e Estatística**, v. 18, p. 83–101, 2000.

LÓPEZ, J. L.; FERREIRA, C. Asymptotic expansions of the Lauricella hypergeometric function f_d. **Journal of Computational and Applied Mathematics**, v. 151, p. 235–256, 2003.

LUGANNANI, R.; RICE, S. Saddlepoint approximation for the distribution of the sum of independent random variables. **Adv. Appl. Prob.**, v. 12, p. 475–490, 1980.

MAFFET, A. L.; WACKERMAN, C. C. The modified beta density function as a model for synthetic aperture radar clutter statistics. **IEEE Transactions on Geoscience and Remote Sensing**, v. 29, p. 277–283, 1991.

MATHAI, A. M.; SAXENA, R. K. Various pratical problems in probability and statistics where lauricella's confluent hypergeometric function appears naturally. **Gadnita Sandesh**, v. 1, n. 1-2, p. 41–48, 1987.

MCCULLAGH, P. **Tensor Methods in Statistics**. [S.l.]: London, 1987.

MCNALLY, R. J. Maximum likelihood estimation of the parameters of the prior distribution of three variables that strongly influence reproductive performance in cows. **Biometrics**, v. 46, n. 2, p. 501–514, 1990.

MILYUTIN, E. R.; YAROMENKO, Y. L. Statistical characteristics of atmospheric transparency index over tilted routes. **Meteorologiya in Gidrologiya**, v. 12, p. 72–76, 1991.

MIYASHIRO, E. S. **Modelos de Regressão Beta e Simplex para Análise de Proporções**. Dissertação (Mestrado) — Instituto de Matemática e Estatística - USP, 2008.

NADARAJAH, S.; GUPTA, A. K. The beta Fréchet distribution. **Far East Journal of Theoretical Statistics**, v. 14, p. 15–24, 2004.

NADARAJAH, S.; KOTZ, S. The beta Gumbel distribution. **Mathematical Problems in Engineering**, v. 4, p. 323–332, 2004.

NADARAJAH, S.; KOTZ, S. The beta exponential distribution. **Reliability Engineering and System Safety**, v. 91, p. 689–697, 2005.

NADARAJAH, S.; KOTZ, S. A generalized beta distribution II. **pre-print**, 2009.

ONG, S. H.; LEE, P. A. Probabilistic interpretations of a transformation of Lauricella's hypergeometric function of n variables and an integral of product of laguerre polynomials. **International J. Math. Statist. Sci.**, v. 9, n. 1, p. 5–13, 2000.

PRUDNIKOV, A. P.; BRYCHKOV, Y. A.; MARICHEV, O. I. **Integrals and Series: Direct Laplace Transforms**. [S.l.]: Gordon and Breach Sciense Publishers, 1992.

REID, N. Saddlepoint methods and statistical inference (with discussion). **Statist. Sci.**, v. 3, p. 213–238, 1988.

ROBINSON, J. Saddlepoint approximations for permutation tests and confidence intervals. **Journal of Royal Statistical Society Serie B**, v. 44, p. 91–101, 1982.

RODRIGUES, K. S. P. **Aperfeiçoamento de Testes de Hipóteses em Modelos de Regressão Não-Lineares Simétricos**. Dissertação (Mestrado) — Universidade Federal de Pernambuco, 2006. Disponível em: <http://www.de.ufpe.br/dissertacao072.pdf>.

RUDIN, W. **Princípios de Análise Matemática**. [S.l.]: Rio de Janeiro: ao livro técnico, 1971. 296 p.

SHARMA, C. K.; SINGH, I. J. some integrals involving the Lauricella functions and the multivariable h-function. **J. Pure Appl. Math.**, v. 21, n. 6, p. 596–604, 1990. Indian.

SRIVASTAVA, H. M. **Some Lauricella Multiple Hypergeometric Series Associated with the Product of Several Bessel Functions**. [S.l.]: Constantin Carathéodory: An International Tribute, 1991. 1304–1341 p.

THOMAS, J. E. et al. **Fundamentos de Engenharia de Petróleo**. [S.l.]: Rio de Janeiro, p. 121, 2001.

VITTINGHOFF, E. et al. **Regression Methods in Biostatistics: linear, logistic, survival, and repeated measures models**. [S.l.]: New York, 2005.

WILEY, J. A.; HERSCHOKORU, S. J.; PADIAU, N. S. Heterogeneity in the probability of HIV transmission per sexual contact: the case of male-to-female transmission in penilevaginal intercourse. **Statistics in Medicine**, v. 8, p. 93–102, 1989.

WINGO, D. R. Fitting three-parameter lognormal models by numerical global optimization. **Computational Statistics and Data Analysis**, v. 2, p. 13–15, 1984.