22º COLÓQUIO BRASILEIRO DE MATEMÁTICA

INTRODUÇÃO À TEORIA ASSINTÓTICA

GAUSS M. CORDEIRO

IMPA 26 - 30 JULHO, 1999

Impa

e deram sugestões úteis. Agradeço à Coordenação do Colóquio Brasileiro de Matemática e, em especial, aos professores Paulo Cordaro (USP) e Jacob Pallis (Diretor do IMPA), pelo convite para escrever este texto. Agradeço ainda ao Oscar P. Silva Neto pelo excelente trabalho de preparação dos originais e aos professores Adiel Almeida (Coordenador do Programa de Pós-Graduação em Engenharia de Produção da UFPE), Carlson Verçosa (Chefe do Departamento de Engenharia Mecânica da UFPE) e Enivaldo Rocha (Chefe do Departamento de Estatística da UFPE) pelas condições oferecidas de apoio a este trabalho.

Finalmente, desejo expressar o meu apreço a minha esposa Zilma Cordeiro pela paciência com o meu isolamento de fins de semana em Gravatá, onde pude escrever este livro.

Rio, abril de 1999

Gauss M. Cordeiro

Índice

Capítulo 1

Fundamentos de Inferência Estatística

1.1 Introdução

A *inferência* é a parte fundamental da Estatística e, claramente, é tão antiga quanto a teoria e os métodos que formam a Estatística atual. As primeiras técnicas de inferência surgiram a mais de 200 anos com os trabalhos de Bayes, DeMoivre, Gauss e Laplace. A inferência estatística baseada diretamente na função de verossimilhança foi proposta por Sir Ronald Fisher em 1912 mas só foi intensificada no período de 1930 a 1940 graças às suas contribuições em problemas de experimentação agrícola.

O processo de inferir a partir dos dados observados sobre parâmetros desconhecidos é parte fundamental da lógica indutiva. A inferência científica se confunde com a inferência estatística quando a conexão entre o "estado da natureza desconhecido" e os fatos observados são expressos em termos probabilísticos, i.e., o mecanismo de geração dos dados é governado por uma componente especificada e um erro estocástico que varia de acordo com uma distribuição de probabilidade (conhecida ou desconhecida). Esta composição define o modelo estatístico que descreve a estrutura probabilística dos dados como função de quantidades de interesse conhecidas e de outros parâmetros possivelmente desconhecidos.

A inferência visa a construir procedimentos ou regras apropriadas de alguma natureza científica baseando-se num certo conjunto de dados, tais como: obter uma estimativa de um parâmetro θ desconhecido, construir um conjunto de valores possíveis de θ que tenha

uma confiabilidade especificada ou decidir sobre um valor previamente concebido para θ. Neste sentido, as atividades fim da inferência são: a estimação, a construção de regiões de confiança e o desenvolvimento de testes de hipóteses.

Várias metodologias de inferência têm sido propostas e as mais importantes são decorrentes das teorias de verossimilhança, Bayesiana, "fiducial" e estrutural. Este texto trata exclusivamente da teoria de verossimilhança. Sobre esta teoria, Sir David R. Cox fez o seguinte comentário: *"The likelihood approach plays a central role in the great majority of statistical theory and it does apply when the main object of the investigation is inferential, i.e., to obtain answers to specific questions about the model."* Na teoria Bayesiana, qualquer incerteza sobre os parâmetros desconhecidos de um modelo estatístico (como por exemplo, a validade do modelo) é expressa em termos de probabilidades que representam "graus de credibilidade" do estatístico Bayesiano. A inferência sobre um parâmetro θ para um certo conjunto de dados é conduzida por uma distribuição a posteriori apropriada para θ. A teoria "fiducial" é certamente a mais difícil e problemática destas teorias, pois alguns dos seus princípios são obscuros e dão origem a interpretações contraditórias. Ela só é considerada relevante quando θ é completamente desconhecido antes da experimentação. Não é necessário supor qualquer distribuição a priori para θ, pois ao aplicá-la obtém-se dos dados uma distribuição de probabilidade para este parâmetro. Finalmente, a teoria estrutural (Fraser, 1968) considera que um experimento tem estrutura própria fora do contexto da família de distribuições proposta para as observações dado θ. Os erros de medição representam características objetivas do processo de geração dos dados e existem independentemente do que foi realmente observado.

Este capítulo aborda os fundamentos da teoria de verossimilhança. Os conceitos básicos de função de verossimilhança, função escore, informação e suficiência são apresentados de forma resumida como pré-requisitos dos Capítulos 4 e 5, onde será discutida a teoria de verossimilhança no contexto de grandes amostras. O leitor poderá consultar o livro de Edwards (1972) para ter uma abordagem ampla das técnicas baseadas na função de verossimilhança.

1.2 Função de verossimilhança

Suponha que y é o valor observado de uma variável aletória $Y = (Y_1, \ldots, Y_n)^T$ caracterizada por uma função de probabilidade ou densidade com forma analítica $f(y;\theta)$ conhecida mas dependente de um vetor $\theta = (\theta_1, \ldots, \theta_p)^T$ de parâmetros desconhecidos. Seja $\Theta \subset \mathbb{R}^p$ o espaço paramétrico representando o conjunto de valores possíveis para o vetor θ. A função $f(y;\theta)$ é denominada *função do modelo estatístico* e define alguma família $\mathcal{F}$ de distribuições de probabilidade. O objetivo da inferência é determinar a distribuição de Y na família $\mathcal{F}$, ou equivalentemente, testar uma hipótese expressa através de θ. A teoria de verossimilhança representa um dos métodos mais comuns de inferência estatística.

A *função de verossimilhança* $L(\theta)$ é definida como sendo igual a função do modelo, embora seja interpretada diferentemente como função de θ para y conhecido. Assim, $L(\theta) = f(y;\theta)$. A *inferência de verossimilhança* pode ser considerada como um processo de obtenção de informação sobre um vetor de parâmetros θ, a partir do ponto y do espaço amostral, através da função de verossimilhança $L(\theta)$. Vários vetores $y's$ podem produzir a mesma verossimilhança ou, equivalentemente, uma dada verossimilhança pode corresponder a um contorno $R(y)$ de vetores amostrais. Este processo produz uma redução de informação sobre θ, disponível em y, que é transferida para as estatísticas suficientes definidas pela função de verossimilhança (vide equação (1.5) a seguir). É impressionante como os conceitos (aparentemente distintos) de *suficiência* e *verossimilhança*, ambos introduzidos por Fisher, estão intimamente relacionados conforme a descrição acima.

A inferência via verossimilhança é fundamentada em princípios genéricos como os descritos a seguir. O *princípio de suficiência* estabelece que vetores de dados distintos com os mesmos valores das estatísticas suficientes para um vetor θ de parâmetros fornecem conclusões idênticas sobre θ. O *princípio fraco de verossimilhança* implica que vetores de dados com verossimilhanças proporcionais produzem as mesmas conclusões sobre θ. Para a validade destes dois princípios, admite-se que o modelo estatístico em investigação é adequado. O *princípio forte de verossimilhança* é relativo a variáveis aleatórias distintas que dependem de um mesmo parâmetro e de um mesmo espaço paramétrico. Supondo que dois modelos são adequados aos vetores de dados y e z em questão, este princípio estabelece que se y e z fornecem verossimilhanças proporcionais, então as conclusões sobre

θ tiradas destes dois vetores de dados são idênticas.

Muito frequentemente, as componentes de Y são mutuamente independentes para todas as distribuições em $\mathcal{F}$ e a verossimilhança de θ reduz-se a

$$L(\theta) = \prod_{i=1}^{n} f(y_i;\theta) \ . \tag{1.1}$$

Usualmente, trabalha-se com a *log-verossimilhança* $\ell(\theta) = \log L(\theta)$, também chamada de *função suporte*. No caso de variáveis aleatórias independentes, a log-verossimilhança é aditiva

$$\ell(\theta) = \sum_{i=1}^{n} \log f(y_i;\theta) \ . \tag{1.2}$$

Em geral, mesmo no caso de variáveis aleatórias dependentes, a log-verossimilhança pode ser dada por uma soma, definindo-a a partir das funções densidade (ou de probabilidade) condicionais. Seja $Y_{(j)} = (Y_1, \ldots, Y_j)^T$ e defina a função densidade condicional de Y_j dado $Y_{(j-1)} = y_{(j-1)}$ por $f_{Y_j|Y_{(j-1)}}(y_j|y_{(j-1)};\theta)$. Assim, a log-verossimilhança de θ é dada por

$$\ell(\theta) = \sum_{j=1}^{n} \log f_{Y_j|Y_{(j-1)}}(y_j|y_{(j-1)};\theta) \ , \tag{1.3}$$

com $Y_{(0)}$ especificando o que for necessário para determinar a distribuição da primeira componente Y_1. A versão (1.3) é importante nos modelos de séries temporais.

Exemplo 1.1 *Suponha que as componentes de Y são geradas por um modelo autoregressivo estacionário de primeira ordem com parâmetro de correlação ρ e média μ, i.e., $Y_j = \mu + \rho(Y_{j-1} - \mu) + \epsilon_j$, onde $\epsilon_2, \ldots, \epsilon_n$ são variáveis aleatórias independentes distribuídas como normal $N(0,\tau)$. A log-verossimilhança (1.3) para $\theta = (\mu, \rho, \tau)^T$ se simplifica pois a distribuição de Y_j dado $Y_{(j-1)} = (Y_1, \ldots, Y_{j-1})^T$ depende somente de Y_{j-1} e contribui para a log-verossimilhança com o termo*

$$\log f_{Y_j|Y_{(j-1)}}(y_j|y_{(j-1)};\theta) = -\frac{1}{2}\log(2\pi\tau) - (2\tau)^{-1}\{y_j - \mu - \rho(y_{j-1} - \mu)\}^2 \ .$$

Assim, a log-verossimilhança total $\ell(\theta)$ *reduz-se a*

$$\ell(\theta) = -\frac{n}{2}\log(2\pi\tau) + \frac{1}{2}\log(1-\rho^2) - (2\tau)^{-1}\{(y_1-\mu)^2 + (y_n-\mu)^2 + (1+\rho^2)\sum_{j=2}^{n-1}(y_j-\mu)^2\} + \frac{\rho}{\tau}\sum_{j=2}^{n}(y_j-\mu)(y_{j-1}-\mu) .$$

A função de verossimilhança informa a ordem natural de preferência entre diversas possibilidades de θ. Um conjunto de dados é mais consistente com um vetor θ do que com outro θ' se a verossimilhança associada a θ for maior do que aquela associada a θ'. Generalizando, entre os possíveis candidatos para estimar o parâmetro verdadeiro θ_0 a partir dos mesmos dados y, o vetor de parâmetros mais plausível é aquele de maior verossimilhança. Neste sentido, o método de máxima verossimilhança (MV) objetiva escolher o valor do vetor θ de parâmetros (ou a hipótese no sentido mais amplo) que fornece a chance mais provável de ocorrer novamente os mesmos dados que ocorreram. Assim, para estimar o vetor verdadeiro θ_0 de parâmetros, escolhe-se aquele vetor de parâmetros que maximiza a função de verossimilhança no espaço paramétrico Θ. Logo, a *estimativa de máxima verossimilhança* (EMV) de θ é o vetor $\hat{\theta}$ que maximiza $L(\theta)$ em Θ, isto é, $L(\hat{\theta}) \geq L(\theta)$ para todo $\theta \in \Theta$. Muitas vezes existe um único vetor de parâmetros que maximiza a verossimilhança em Θ, sendo portanto o único vetor mais plausível neste espaço paramétrico. Entretanto, a EMV pode não ser única e nem mesmo finita dentro de um dado espaço de parâmetros. A EMV $\hat{\theta}$ desempenha um papel central na inferência paramétrica em grandes amostras (vide Capítulo 4).

Como a função logaritmo é monótona, maximizar $L(\theta)$ e $\ell(\theta)$ em Θ são processos equivalentes. Então, a EMV $\hat{\theta}$ é definida de modo que para todo $\theta \in \Theta$

$$l(\hat{\theta}) \geq \ell(\theta) . \tag{1.4}$$

O gráfico de $\ell(\theta)$ versus θ em Θ é chamado superfície suporte. Para $p = 1$ este gráfico (curva suporte) é bastante informativo, embora não tenha valor imediato no cálculo de $\hat{\theta}$. Para $p \geq 3$ a superfície suporte não pode ser traçada e deve-se recorrer a técnicas iterativas apresentadas na Seção 1.4. Se Θ é um conjunto discreto, computa-se $\ell(\theta)$ para os diversos

$\theta's$ e escolhe-se $\hat{\theta}$ como aquele valor de θ correspondente ao máximo $\ell(\theta)$. Quando $\ell(\theta)$ é contínua e diferenciável em Θ, a EMV $\hat{\theta}$ pode ser obtida resolvendo-se o sistema de equações simultâneas $\partial\ell(\theta)/\partial\theta_r = 0$ para $r = 1, \ldots, p$ desde que θ não se encontre na fronteira do espaço paramétrico. Das soluções deste sistema (em geral não-linear) pode-se achar a EMV $\hat{\theta}$. Convém frisar, entretanto, que a EMV não coincide necessariamente com alguma solução do sistema. Mesmo que o sistema tenha solução única, não significa que ela seja a EMV, que pode até mesmo nem existir.

Como foi enfatizado anteriormente, a função de verossimilhança resume toda a informação relevante sobre um vetor de parâmetros e, em especial, o quociente de verossimilhanças ou a diferença entre log-verossimilhanças expressa as plausibilidades relativas de dois vetores de parâmetros especificados. Assim, a verossimilhança retira dos dados toda a informação relevante para inferir sobre um vetor de parâmetros de interesse e a sua "inspeção" possibilita responder questões específicas sobre estes parâmetros. Toda informação relevante na verossimilhança sobre um vetor de parâmetros está contida num conjunto de estatísticas denominadas suficientes, definidas a seguir. Assim, um conceito diretamente relacionado à verossimilhança é a *suficiência*. Considere-se uma estatística $S = S(Y)$ função das variáveis aleatórias $Y_1, \ldots, Y_n$. Seja s o valor observado de S. Diz-se que S é suficiente para θ na família de distribuições definida por $\mathcal{F}$ se a distribuição condicional $f(y|s)$ de $Y = (Y_1, \ldots, Y_n)^T$ dado $S = s$ independe de θ. A suficiência de S implica que toda informação relevante que os dados y contêm sobre θ está concentrada em S. Uma condição necessária e suficiente para esta suficiência é que a verossimilhança possa ser fatorada na forma

$$L(\theta) = g(s, \theta)h(y) \ , \tag{1.5}$$

onde $g(\cdot, \cdot)$ depende dos dados y somente através de $s = s(y)$ e $h(y)$ é uma função dos dados que independe de θ. A condição (1.5) é conhecida como o *Teorema da Fatoração de Neyman-Fisher*. Uma demonstração detalhada (o caso discreto é trivial) pode ser encontrada no livro de Lehmann (1959, p.470). Claro que se S é suficiente para θ, qualquer função um-a-um de S também é suficiente. A escolha entre distintas estatísticas suficientes para um parâmetro pode ser baseada na consistência, eficiência e no fato de ser não-viesada (Seção 4.1.1).

Uma propriedade que relaciona a suficiência e a verossimilhança pode ser deduzida diretamente da fatoração (1.5). Se existe um conjunto de estatísticas $S_1, \ldots, S_m$ conjuntamente suficientes para os parâmetros $\theta_1, \ldots, \theta_p$, segue-se de (1.5) que maximizar $L(\theta)$ equivale a maximizar a distribuição conjunta dessas estatísticas (identificada como $g(s, \theta)$) em relação aos parâmetros. Então, as estimativas de MV $\hat{\theta}_1, \ldots, \hat{\theta}_p$ devem ser funções de $S_1, \ldots, S_m$. Entretanto, as dimensões m e p de S e θ, respectivamente, não são necessariamente iguais. O caso $m < p$ poderá ocorrer se existirem relações não-lineares entre as componentes de θ, mas a situação mais comum na prática é $m \geq p$. Como as componentes do vetor $\hat{\theta}$ podem não ser funções um a um das estatísticas suficientes $S_1, \ldots, S_m$, as estimativas $\hat{\theta}_1, \ldots, \hat{\theta}_p$ não formam necessariamente um conjunto de estatísticas suficientes para θ, pois podem ser apenas funções de um subconjunto dessas estatísticas.

Usando-se a definição de suficiência ou a condição (1.5) é fácil mostrar, por exemplo, que no caso de observações *iid* (independentes e identicamente distribuídas), a média amostral é suficiente para a média da distribuição de Poisson e para a probabilidade de sucesso da distribuição binomial. Pode-se ainda verificar no caso *iid* que se $Y \sim N(\mu, \sigma^2)$ a verossimilhança para $\theta = (\mu, \sigma^2)^T$ pode ser fatorada como (1.5) com $g(\overline{y}, s^2, \mu, \sigma^2)$ onde $\overline{y} = \Sigma y_i/n$ e $s^2 = \Sigma(y_i - \overline{y})^2/n$ e, portanto, a média $\overline{y}$ e a variância s^2 amostrais são estatísticas conjuntamente suficientes para μ e σ^2. Entretanto, s^2 sozinha não será suficiente para σ^2 quando μ for desconhecido. A partir da log-verossimilhança do modelo autoregressivo discutido no exemplo 1.1, observa-se que as estatísticas $y_1^2 + y_n^2$, $\sum_{j=2}^{n-1} y_j^2$ e $\sum_{j=2}^{n} y_j y_{j-1}$ são suficientes para os parâmetros ρ e τ quando μ é conhecido.

A inferência através da função suporte deve ser consistente com os dados observados e, portanto, as conclusões não deverão ser alteradas por dois tipos de transformações: (i) transformação inversível de Y; (ii) transformação não necessariamente inversível de θ.

Mostra-se agora que a função suporte quando usada relativamente é invariante segundo transformação unívoca dos dados. Supondo uma transformação um-a-um da variável aleatória contínua Y para $Z = Z(Y)$, a verossimilhança segundo os novos dados $z\,(L^*(\theta; z))$ pode ser expressa em termos da verossimilhança segundo os dados $y\,(L(\theta; y))$ por

$$L^*(\theta; z) = L(\theta; y)|T|\,, \tag{1.6}$$

onde $T = \frac{\partial y}{\partial z}$ é o Jacobiano da transformação de Y para Z suposto não-nulo. De (1.6) vem $\ell^*(\theta; z) = \ell(\theta; y) + \log |T|$, o que demonstra a invariância da função suporte em relação à transformação dos dados.

A função suporte relativa a um novo parâmetro ϕ, supondo que os dados são mantidos constantes, onde $\phi = f(\theta)$ e f é uma transformação um-a-um, é encontrada diretamente substituindo θ por $f^{-1}(\phi)$. Tem-se $\ell(\theta) = \ell(f^{-1}(\phi)) = \ell^*(\phi)$, onde ℓ e ℓ^* são os suportes em termos de θ e ϕ, respectivamente. Se $\hat{\theta}$ é a EMV de θ, obtém-se $\ell(\hat{\theta}) \geq \ell(\theta)$ para qualquer θ. Definindo $\hat{\phi} = f(\hat{\theta})$ vem, para todo ϕ, $\ell(f^{-1}(\hat{\phi})) \geq \ell(f^{-1}(\phi))$ ou seja $\ell^*(\hat{\phi}) \geq \ell^*(\phi)$, o que implica $\hat{\phi}$ ser a EMV de $\phi = f(\theta)$. Note-se que as superfícies suportes $\ell(\theta)$ e $\ell^*(\phi)$ têm formas distintas, porém o mesmo máximo $\ell(\hat{\theta}) = \ell^*(\hat{\phi})$. Assim, o valor da verossimilhança maximizada segundo um modelo estatístico é único, qualquer que seja a parametrização adotada para o modelo. A propriedade de invariância estabelece que a EMV de $f(\theta)$ é a função f avaliada na EMV de θ. Ela é importante, pois alguma parametrização do modelo pode conduzir a simplificações mais consideráveis no cálculo da EMV. A demonstração desta propriedade é imediata usando a regra da cadeia no caso de $f(\theta)$ ser diferenciável.

1.3 Função Escore e Informação

A primeira derivada da função suporte é chamada *função* (ou vetor) *escore*

$$U(\theta) = \frac{\partial \ell(\theta)}{\partial \theta}\,, \tag{1.7}$$

onde o operador $\frac{\partial}{\partial \theta}$ é interpretado como um vetor coluna e, portanto, $U(\theta)$ é um vetor $p \times 1$. Assim, $U(\theta)$ é o vetor gradiente da superfície suporte em θ. As equações de MV são expressas por $U(\hat{\theta}) = 0$ mostrando que a função escore é zero em $\hat{\theta}$.

As equações de MV são usualmente não-lineares e nestes casos as soluções de $U(\hat{\theta}) = 0$ devem ser obtidas por técnicas iterativas. Quando as EMV têm forma fechada, pode ser viável determinar suas distribuições exatas e, portanto, obter suas propriedades em pequenas amostras. Quando este não é o caso, a inferência deve ser baseada na teoria assintótica apresentada nos Capítulos 4 e 5.

Como ilustração do cálculo de EMV, considere n observações *iid* da distribuição nor-

mal $N(\mu,\sigma^2)$ e da distribuição de Cauchy, cuja densidade é $f(y;\theta) = \pi^{-1}\{1+(y-\theta)^2\}^{-1}$, $y \in \mathbb{R}$, com o parâmetro θ representando a mediana da distribuição. No caso da normal, as EMV são facilmente obtidas de $\hat{\mu} = \overline{y}$ e $\hat{\sigma}^2 = s^2$, i.e., igualam as estatísticas conjuntamente suficientes para estes parâmetros. Sabe-se que $\hat{\mu} \sim N(\mu, \sigma^2/n)$ e $\hat{\sigma}^2 \sim \frac{\sigma^2}{n}\chi^2_{n-1}$ e como suas distribuições são independentes, $\sqrt{n-1}(\overline{y}-\mu)/s$ tem distribuição t_{n-1} (t de Student com $n-1$ graus de liberdade). Estes resultados possibilitam determinar intervalos de confiança exatos para os parâmetros da normal ou de qualquer distribuição definida por uma transformação a partir da distribuição normal. A idéia de transformar uma variável de modo a obter normalidade é de grande interesse na Estatística. Por exemplo, se $Y \sim N(\mu,\sigma^2)$ define-se a distribuição *lognormal* ($Z \sim LN(\mu,\sigma^2)$) de dois parâmetros por $Z = \exp(Y)$. É evidente que a estimação por MV dos parâmetros em qualquer parametrização de Z é feita através das estimativas $\hat{\mu}$ e $\hat{\sigma}^2$. Por exemplo, a EMV do r-ésimo momento $\mu'_r = E(Z^r)$ de Z é simplesmente $\hat{\mu}'_r = \exp(r\hat{\mu} + r^2\hat{\sigma}^2/2)$ para $r \geq 1$. No caso da estimação do parâmetro θ da distribuição de Cauchy (exemplo 1.4 dado a seguir), a equação de MV não tem forma simples, sendo representada por um polinômio de grau $n-1$ em θ cujas soluções em geral incluem vários máximos e mínimos da log-verossimilhança. Portanto, a inferência sobre θ deve ser baseada em propriedades assintóticas de sua EMV $\hat{\theta}$.

A *matriz de informação* (algumas vezes chamada *informação esperada*) para $\theta \in \mathbb{R}^p$ obtida dos dados y é uma matriz $p \times p$ definida por

$$K(\theta) = E\{U(\theta)U(\theta)^T\}\ . \tag{1.8}$$

Para observações independentes, a função escore e a informação são somas de contribuições individuais sobre θ.

Este texto considera apenas problemas regulares que satisfazem às seguintes condições: (a) Θ é fechado, compacto e tem dimensão finita sendo o parâmetro verdadeiro θ_0 um ponto interior de Θ; (b) $f(y;\theta)$ é uma função um-a-um de θ; (c) as três primeiras derivadas de $\ell(\theta)$ existem numa vizinhança de θ_0; (d) $K(\theta)$ é finita e positiva definida numa vizinhança de θ_0. Além das condições (a)-(d), admite-se, para modelos contínuos,

que a igualdade

$$\frac{\partial}{\partial\theta}E\{t(Y)\} = \int t(y)\frac{\partial}{\partial\theta}f(y;\theta)dy$$

é válida para qualquer estatística $t(Y)$. Para modelos discretos basta substituir esta integral por um somatório. Esta equação garante que as operações de diferenciação com respeito a θ e integração em y são permutáveis. Isso é possível, por exemplo, se os limites de variação de y são finitos e independem de θ ou, no caso de infinitos, se a integral resultante da permutação é convergente para todo θ e o integrando é uma função contínua de y e θ. Estas condições de regularidade serão rediscutidas na Seção 4.1.3.

As condições anteriores são usadas para justificar expansões em séries de Taylor e técnicas similares. Uma discussão mais detalhada destas condições pode ser encontrada em LeCam (1956, 1970). De agora em diante omite-se o argumento θ das funções de verossimilhança, suporte, escore e informação, escrevendo abreviadamente estas quantidades como L, ℓ, U e K. Ainda, a distribuição conjunta dos dados é escrita apenas como f sem os argumentos y e θ. As demonstrações serão dadas em forma resumida para modelos contínuos. Para modelos discretos, basta substituir a integral por um somatório.

A esperança e a covariância da função escore são dadas por

$$E(U) = 0 \tag{1.9}$$

e

$$\mathrm{Cov}(U) = E\left(-\frac{\partial U^T}{\partial\theta}\right) = E\left(-\frac{\partial^2\ell}{\partial\theta\partial\theta^T}\right) = K, \tag{1.10}$$

respectivamente. De (1.7) $U = \frac{1}{f}\frac{\partial f}{\partial\theta}$ e, então, $E(U) = \int \frac{\partial f}{\partial\theta}dy = \frac{\partial}{\partial\theta}(\int f dy) = 0$. Diferenciando $\int Uf dy = 0$ em relação a θ vem $\int\{\frac{\partial U^T}{\partial\theta}f + U\frac{\partial f^T}{\partial\theta}\}dy = \int\{\frac{\partial U^T}{\partial\theta} + UU^T\}f dy = 0$. Por (1.8) e (1.9) obtém-se (1.10). Esta equação implica que o elemento (r,s) de K pode ser calculado de duas formas, como $-E\{\frac{\partial^2\ell}{\partial\theta_r\partial\theta_s}\}$ ou $E\{\frac{\partial\ell}{\partial\theta_r}\frac{\partial\ell}{\partial\theta_s}\}$, sendo a primeira geralmente mais fácil. De agora em diante, quantidades avaliadas na EMV $\hat{\theta}$ serão escritas com superescritos $\wedge$.

A matriz de primeiras derivadas da função escore com sinal negativo $J = -\frac{\partial U^T}{\partial\theta} = -\frac{\partial^2\ell}{\partial\theta\partial\theta^T}$ é denominada matriz de *informação observada*. A matriz Hessiana é simplesmente $-J$ e tem-se $E(J) = K$. Para $\hat{\theta}$ ser um máximo local, as condições $\hat{U} = 0$ e $\hat{J} \geq 0$ ($\hat{J}$

positiva semi-definida) são *necessárias* enquanto que $\hat{U} = 0$ e $\hat{J} > 0$ ($\hat{J}$ positiva definida) são *suficientes.*

Exemplo 1.2 *Se* $Y = (Y_1, \ldots, Y_n)^T$ *e os* $Y_i's$ *são variáveis aleatórias iid tendo distribuição exponencial com função densidade* $\rho e^{-\rho y}$, *então a log-verossimilhança e a função escore para* ρ *são, respectivamente,* $\ell(\rho) = n \log \rho - \rho \sum_{i=1}^{n} y_i$ *e* $U(\rho) = n/\rho - \sum_{i=1}^{n} y_i$. *É simples checar diretamente que* $E\{U(\rho)\} = 0$ e $\mathrm{Var}\{U(\rho)\} = n/\rho^2$.

Exemplo 1.3 *A função de probabilidade em série de potências* $SP(\theta)$ *é definida por* $P(Y = y; \theta) = a_y \theta^y / f(\theta)$ *para* $y = 0, 1, \ldots$ *e* $\theta > 0$, *onde* $a_y \geq 0$ *e* $f(\theta) = \sum_{y=0}^{\infty} a_y \theta^y$. *Supondo que as observações são iid, a função de verossimilhança é expressa por* $L(\theta) = \theta^{n\overline{y}} f(\theta)^{-n} \prod_{i=1}^{n} a_{y_i}$, *sendo* $\overline{y}$ *a média amostral. A EMV* $\hat{\theta}$ *é uma função não-linear de* $\overline{y}$ *obtida iterativamente de* $\overline{y}/\hat{\theta} - f'(\hat{\theta})/f(\hat{\theta}) = 0$. *A média amostral* $\overline{y}$ *é suficiente para* θ *e a informação para* θ *é dada por*

$$K(\theta) = \frac{n}{\theta f(\theta)}[f'(\theta) + \theta\{f(\theta)f'(\theta) - f'(\theta)^2\}] \, .$$

Expandindo o suporte ℓ em θ em série multivariada de Taylor ao redor de $\hat{\theta}$ e notando que $\hat{U} = 0$ obtém-se, aproximadamente,

$$\hat{\ell} - \ell = \frac{1}{2}(\theta - \hat{\theta})^T \hat{J} (\theta - \hat{\theta}) \, . \tag{1.11}$$

A equação (1.11) revela que a diferença entre o máximo suporte e o suporte num ponto arbitrário, que pode ser vista como a quantidade de informação dos dados sobre θ, é proporcional a $\hat{J}$ (i.e. à informação observada no ponto $\hat{\theta}$). O determinante de $\hat{J}(|\hat{J}|)$ pode ser interpretado geometricamente como a curvatura esférica da superfície suporte no seu ponto máximo. A forma quadrática do lado direito de (1.11) aproxima a superfície suporte por um parabolóide, passando pelo seu ponto de máximo, com a mesma curvatura esférica da superfície neste ponto. O recíproco de $|\hat{J}|$ mede a variabilidade de θ ao redor da EMV $\hat{\theta}$. E, como esperado, quanto maior a informação sobre θ, menor será a dispersão de θ ao redor de $\hat{\theta}$.

A interpretação geométrica dos conceitos acima é melhor compreendida no caso uniparamétrico, onde (1.11) reduz-se a equação de uma parábola $\ell = \hat{\ell} - \frac{1}{2}(\theta - \hat{\theta})^2 \hat{J}$. Uma inspeção gráfica mostra que esta parábola aproxima a curva suporte, coincidindo no seu ponto máximo e tendo a mesma curvatura desta curva em $\hat{\theta}$, revelando ainda que quanto maior a curvatura menor a variação de θ em torno de $\hat{\theta}$.

A equação (1.11) implica que a verossimilhança L num ponto qualquer θ segue, aproximadamente, a expressão

$$L = \hat{L} \exp\left\{-\frac{1}{2}(\theta - \hat{\theta})^T \hat{J}(\theta - \hat{\theta})\right\} , \tag{1.12}$$

que representa a forma de curva normal multivariada com média $\hat{\theta}$ e estrutura de covariância igual a $\hat{J}^{-1}$. Através desta aproximação pode-se então tratar o vetor de parâmetros como se fosse um vetor de variáveis aleatórias tendo distribuição normal multivariada com média igual à EMV $\hat{\theta}$ e estrutura de covariância $\hat{J}^{-1}$. Quando o suporte for *quadrático*, a verossimilhança terá a forma *normal*. A forma de L se aproximará cada vez mais da distribuição normal quando n tender para infinito.

A fórmula (1.12) mostra a fatoração da verossimilhança como (1.5) pelo menos para n grande, estabecelendo a *suficiência assintótica* da EMV (Seção 4.2). Conclui-se que, embora as EMV não sejam necessariamente suficientes para os parâmetros do modelo, esta suficiência será alcançada quando a dimensão do vetor de dados tender para infinito.

Convém citar nesta seção algumas propriedades da matriz de informação. Seja $K_y(\theta)$ a informação sobre um vetor paramétrico θ contida nos dados y obtidos de certo experimento. A informação é aditiva para amostras y e z independentes, isto é, $K_{y+z}(\theta) = K_y(\theta) + K_z(\theta)$. Esta igualdade implica que a informação contida numa amostra de tamanho n de observações *iid* é igual a n vezes a informação devida a uma única observação. Como seria previsto, a informação (esperada ou observada) sobre θ contida nos dados mantém-se invariante segundo qualquer transformação um-a-um desses dados. Como conseqüência direta de (1.6), obtém-se $K_z(\theta) = K_y(\theta)$ se $z = z(y)$. Uma propriedade procedente do teorema da fatoração expressa que a informação sobre θ fornecida por uma estatística suficiente $s = s(y)$ é a mesma daquela fornecida pelos dados y. Em símbolos, $K_s(\theta) = K_y(\theta)$.

Em geral, para qualquer estatística $t = t(y)$ definida pela sua função de probabilidade ou função densidade $g_t(x;\theta)$ tem-se $K_t(\theta) \leq K_y(\theta)$. A igualdade ocorrerá se e somente se t for suficiente para θ. Para demonstrar esta importante desigualdade basta desenvolver $E[\{U(\theta) - \frac{\partial}{\partial\theta}\log g_t(x;\theta)\}^2]$ e usar a fórmula da esperança condicional da função escore dado $t = x$, ou seja,

$$E\{U(\theta)|t = x\} = \frac{\partial}{\partial\theta}\log g_t(x;\theta) .$$

Assim, a redução de uma amostra por uma estatística poderá implicar perda de informação relativa a um parâmetro desconhecido. Entretanto, não haverá perda se e somente se a suficiência for mantida no processo de redução dos dados.

As propriedades da EMV e alguns critérios para a estimação paramétrica serão discutidos na Seção 4.1.

1.4 Métodos Iterativos

Os métodos iterativos para o cálculo da EMV são bastante utilizados na prática e, em geral, mostram-se imprescindíveis quando a dimensão p do espaço de parâmetros é grande. Expandindo $\hat{U}$ (a função escore em $\hat{\theta}$) em série multivariada de Taylor até primeira ordem ao redor de um ponto qualquer θ pertencente a uma vizinhança de $\hat{\theta}$, tem-se, aproximadamente,

$$\hat{U} = U + \frac{\partial U^T}{\partial\theta}(\theta - \hat{\theta}) .$$

Como $\hat{U} = 0$ obtém-se a relação aproximada

$$\hat{\theta} - \theta = J^{-1}U \tag{1.13}$$

entre a EMV e a função escore e a informação observada avaliadas no ponto θ próximo de $\hat{\theta}$. O método de Newton-Raphson para o cálculo da EMV consiste em usar a equação (1.13) iterativamente. Obtém-se uma nova estimativa $\theta^{(m+1)}$ a partir de uma anterior $\theta^{(m)}$ através de

$$\theta^{(m+1)} = \theta^{(m)} + J^{(m)^{-1}}U^{(m)} , \tag{1.14}$$

onde quantidades avaliadas na m-ésima iteração do procedimento iterativo são indicadas com o superescrito (m). O processo é então repetido até a distância entre $\theta^{(m+1)}$ e $\theta^{(m)}$ se

tornar desprezível ou menor que uma quantidade pequena especificada. Geometricamente, uma iteração do método equivale a ajustar um parabolóide à superfície suporte em $\theta^{(m)}$, tendo o mesmo gradiente e curvatura da superfície neste ponto, e então obter o ponto máximo do parabolóide que corresponderá à estimativa atualizada $\theta^{(m+1)}$. Quando θ é um escalar, a equação (1.14) reduz-se a $\theta^{(m+1)} = \theta^{(m)} - U^{(m)}/U'^{(m)}$, onde $U' = \frac{dU}{d\theta}$, que representa o método das tangentes, bastante usado para calcular a solução de uma equação não-linear $\hat{U} = 0$.

A seqüência $\{\theta^{(m)}; m > 1\}$ gerada depende fundamentalmente do vetor inicial $\theta^{(1)}$, dos valores amostrais e do modelo estatístico e, em determinadas situações, onde n é pequeno, pode revelar irregularidades específicas aos valores amostrais obtidos do experimento e, portanto, pode não convergir e mesmo divergir da EMV $\hat{\theta}$. Mesmo existindo a convergência, se a verossimilhança tem raízes múltiplas, não há garantia de que o procedimento converge para a raiz correspondente ao maior valor absoluto da verossimilhança. No caso uniparamétrico, se a estimativa inicial $\theta^{(1)}$ for escolhida próxima de $\hat{\theta}$ e se $J^{(m)}$ para $m \geq 1$ for limitada por um número real positivo, existirá uma chance apreciável que esta seqüência vá convergir para $\hat{\theta}$.

A expressão (1.13) tem uma forma alternativa assintótica equivalente, pois pela lei dos grandes números J deve convergir para K quando $n \to \infty$ (vide Seção 4.1.4). Assim, substituindo a informação observada em (1.13) pela esperada, obtém-se a aproximação

$$\hat{\theta} - \theta = K^{-1}U \ . \tag{1.15}$$

O procedimento iterativo baseado em (1.15) é denominado *método escore* de Fisher para parâmetros, i.e., $\theta^{(m+1)} = \theta^{(m)} + K^{(m)^{-1}}U^{(m)}$. O aspecto mais trabalhoso dos dois esquemas iterativos é a inversão das matrizes J e K. Ambos os procedimentos são muitos sensíveis em relação à estimativa inicial $\theta^{(1)}$. Se o vetor $\theta^{(1)}$ for uma estimativa consistente, os métodos convergirão em apenas um passo para uma estimativa eficiente assintoticamente (Seção 4.1.7).

Existe evidência empírica que o método de Fisher é melhor, em termos de convergência, do que o método de Newton-Raphson. Ela possui ainda a vantagem de usufruir

(através da matriz de informação) de características específicas ao modelo estatístico. Ademais, em muitas situações, é mais fácil determinar a inversa de K em forma fechada do que a inversa de J, sendo a primeira menos sensível a variações em θ do que a segunda. Neste sentido, K pode ser considerada aproximadamente constante em todo o processo iterativo, requerendo que a inversão seja feita apenas uma vez. Uma vantagem adicional do método escore é que usa-se a matriz K^{-1} para obter aproximações de primeira ordem para as variâncias e covariâncias das estimativas $\hat{\theta}_1, \ldots, \hat{\theta}_p$ como será visto na Seção 4.1.6.

Exemplo 1.4 *No caso da função densidade de Cauchy* $f(y;\theta) = \pi^{-1}\{1 + (y-\theta)^2\}^{-1}$, *apresentada na Seção 1.3, mostra-se facilmente que a informação é* $K = \{\frac{n}{2}\}$ *e o processo iterativo (1.14) segue de*

$$\theta^{(m+1)} = \theta^{(m)} + \frac{4}{n}\sum_{i=1}^{n} \frac{y - \theta^{(m)}}{1 + (y_i - \theta^{(m)})^2} \cdot$$

Exemplo 1.5 *A função densidade de Weibull* $W(\alpha, \phi)$ *é dada por*

$$f(y;\alpha,\phi) = \frac{\alpha}{\phi}\left(\frac{y}{\phi}\right)^{\alpha-1} \exp\left\{-\left(\frac{y}{\phi}\right)^{\alpha}\right\}$$

com $\alpha > 0$ *e* $\phi > 0$. *Supondo observações iid, as EMV são expressas por*

$$\hat{\alpha} = \left(\frac{\sum_i y_i^{\hat{\alpha}} \log y_i}{\sum_i y_i^{\hat{\alpha}}} - \log \tilde{y}\right)^{-1} \tag{1.16}$$

e

$$\hat{\phi} = \left(n^{-1}\sum_i y_i^{\hat{\alpha}}\right)^{1/2}, \tag{1.17}$$

onde $\tilde{y}$ *é a média geométrica dos dados. A EMV* $\hat{\alpha}$ *é calculada iterativamente de (1.16) e depois obtém-se* $\hat{\phi}$ *de (1.17). A matriz de informação de* α *e* ϕ *é dada por*

$$K = \begin{array}{c} \\ \alpha \\ \phi \end{array} \begin{array}{c} \begin{array}{cc} \alpha & \phi \end{array} \\ \begin{pmatrix} \frac{\pi^2/6 + \Gamma'(2)^2}{\alpha^2} & -\frac{\Gamma'(2)}{\phi} \\ -\frac{\Gamma'(2)}{\phi} & \frac{\alpha^2}{\phi^2} \end{pmatrix} \end{array},$$

onde $\Gamma(p) = \int_0^\infty x^{p-1}e^{-x}dx$ *é a função gama e* $\Gamma'(p)$ *a sua derivada.*

1.5 Modelos Exponenciais

Suponha que p parâmetros desconhecidos $\theta = (\theta_1, \ldots, \theta_p)^T$ e p estatísticas (i.e. funções dos dados y) $s = (s_1, \ldots, s_p)^T$ são tais que a função densidade (ou de probabilidade no caso discreto) de $Y = (Y_1, \ldots, Y_n)^T$ possa ser expressa como

$$f(y;\theta) = h(y)\exp\{s^T\theta - b(\theta)\}\ , \tag{1.18}$$

onde as componentes de $s = s(y)$ são linearmente independentes. O modelo (1.18) é denominado *modelo exponencial* com parâmetros canônicos $\theta_1, \ldots, \theta_p$ e estatísticas suficientes $s_1, \ldots, s_p$. Observa-se que (1.18) tem a forma (1.5). O espaço paramétrico Θ consiste de todos os $\theta's$ tais que $\int h(y)\exp(s^T\theta)dy < \infty$. A quantidade $\exp\{-b(\theta)\}$ representa a constante normalizadora de modo a tornar a integral (1.18) igual a 1.

O modelo exponencial (1.18) é de grande interesse pois inclui várias distribuições importantes na análise de dados, tais como, normal, gama, Poisson e binomial, como casos especiais. Cordeiro, Ferrari, Aubin e Cribari-Neto (1996) listam 24 distribuições importantes no modelo exponencial uniparamétrico ($p = 1$).

Exemplo 1.6 *Considere o modelo de regressão normal linear* $Y \sim N(\mu, \sigma^2 I)$, *onde* $\mu = E(Y) = X\beta$ *e* X *é uma matriz* $n \times p$ *conhecida,* $\beta \in \mathbb{R}^p$ *é um vetor de parâmetros desconhecidos e* σ^2 *é a variância comum desconhecida. A log-verossimilhança para os parâmetros* $\theta = (\beta^T, \sigma^2)^T$ *pode ser escrita como*

$$\ell(\beta, \sigma^2) = -\frac{n}{2}\log\,\sigma^2 - \frac{1}{2\sigma^2}(y - X\beta)^T(y - X\beta)\ . \tag{1.19}$$

Maximizando (1.19) obtêm-se as EMV $\hat{\beta} = (X^TX)^{-1}X^Ty$ *e* $\hat{\sigma}^2 = SQR/n$, *onde* $SQR = (y - X\hat{\beta})^T(y - X\hat{\beta})$. *A forma da log-verossimilhança para o modelo normal mostra que a EMV de* β *iguala aquela de mínimos quadrados correspondente à minimização de* $(y - X\beta)^T(y - X\beta)$. *A forma explícita de* $\hat{\beta}$ *implica*

$$(y - X\beta)^T(y - X\beta) = (y - X\hat{\beta})^T(y - X\hat{\beta}) + (\hat{\beta} - \beta)^TX^TX(\hat{\beta} - \beta)\ .$$

Assim, os dados y entram na log-verossimilhança (1.19) através das estimativas $\hat{\beta}$ e da soma de quadrados dos resíduos SQR. Então, as estatísticas suficientes para $(\beta^T, \sigma^2)^T$ são $(\hat{\beta}^T, SQR)^T$. Quando σ^2 é conhecido, $\hat{\beta}$ é a estatística suficiente para β.

Observe-se que o modelo normal linear pertence à família exponencial (1.18) pois a verossimilhança pode ser expressa por

$$L(\theta) = f(y;\theta) = \frac{1}{(2\pi)^{n/2}} \exp\left\{ y^T y \left(-\frac{1}{2\sigma^2}\right) + \hat{\beta}^T \left(\frac{(X^TX)^{-1}\beta}{\sigma^2}\right) - \frac{\beta^T (X^TX)^{-1}\beta}{2\sigma^2} - \frac{n}{2}\log \sigma^2 \right\},$$

sendo as estatísticas suficientes $(\hat{\beta}^T, y^T y)$. Este exemplo ilustra que a suficiência é preservada segundo transformação um-a-um, pois $y^T y = SQR + \hat{\beta}^T (X^TX)^{-1}\hat{\beta}$.

A função escore e a informação para o modelo (1.18) são obtidas de (1.7) e (1.8), respectivamente, como

$$U(\theta) = s - \frac{\partial b(\theta)}{\partial \theta} \quad \text{e} \quad K(\theta) = \frac{\partial^2 b(\theta)}{\partial \theta \partial \theta^T}.$$

Usando (1.9) verifica-se que o vetor S de estatísticas suficientes tem esperança $E(S) = \partial b(\theta)/\partial\theta$. Além disso, obtém-se de (1.10) a matriz $(p \times p)$ de covariância de S como $\mathrm{Cov}(S) = \partial^2 b(\theta)/\partial\theta\partial\theta^T$. No exemplo 2.5 (Seção 2.3) mostra-se que $b(\cdot)$ em (1.18) é a função geradora de cumulantes de S e, portanto, os casos acima se referem aos dois primeiros cumulantes de S.

A EMV $\hat{\theta}$ do parâmetro canônico θ em modelos exponenciais é solução da equação

$$\left.\frac{\partial b(\theta)}{\partial \theta}\right|_{\hat{\theta}} = s,$$

ou seja, é obtida igualando $E(S)$ avaliado em $\hat{\theta}$ ao valor observado s do vetor S de estatísticas suficientes.

1.6 Estimação por Intervalos

Suponha que Y tem função densidade ou função de probabilidade $f(y;\theta)$ dependendo de um parâmetro real θ desconhecido. A partir dos dados y constroem-se *intervalos de confiança* para θ através de uma *quantidade pivotal* $\rho(t,\theta)$ cuja distribuição pode ser obtida (pelo menos aproximadamente) não dependendo de θ, onde $t = t(y)$ é uma estimativa pontual razoável de θ. Da distribuição de $\rho(t,\theta)$ calculam-se os limites a e b tais que

$$P(a \leq \rho(t,\theta) \leq b) = 1 - \alpha \,, \tag{1.20}$$

onde $1 - \alpha$ é uma confiabilidade especificada. Suponha ainda que, para t fixo, $\rho(t,\theta)$ seja uma função monótona de θ. Então, observado t, a desigualdade em (1.20) pode ser invertida para produzir uma região de valores de θ com confiabilidade $1 - \alpha$. Esta região é frequentemente um intervalo do tipo

$$P\{k_1(t) \leq \theta \leq k_2(t)\} = 1 - \alpha \,, \tag{1.21}$$

onde $k_1(t)$ e $k_2(t)$ são funções de t, a e b mas não envolvem θ. O conjunto $[k_1(t), k_2(t)]$ é um intervalo de $100(1 - \alpha)\%$ de confiança para θ. A generalização para um vetor θ será feita nas Seções 4.3 e 4.4. A desigualdade em (1.21) deve ser cuidadosamente interpretada. Como os limites em (1.21) são aleatórios, não se pode interpretar $1 - \alpha$ como a probabilidade do parâmetro verdadeiro θ_0 estar em algum intervalo observado. Isto só teria sentido se o parâmetro desconhecido fosse uma variável aleatória e os limites $k_1(t)$ e $k_2(t)$ constantes. Contrariamente, os intervalos do tipo $[k_1(t), k_2(t)]$ serão em geral diferentes para amostras diferentes. Alguns deles conterão o valor verdadeiro de θ enquanto outros não. Assim, deve-se interpretar $1 - \alpha$ como a freqüência esperada dos casos, numa longa série de amostras independentes, em que os intervalos $[k_1(t), k_2(t)]$ conterão θ_0.

A distribuição assintótica $N(\theta, K(\theta)^{-1})$ da EMV $\hat{\theta}$ do escalar θ (Seção 4.1.6) possibilita construir um intervalo aproximado para este parâmetro, supondo que $(\hat{\theta} - \theta)K(\hat{\theta})^{-1/2}$ tem distribuição $N(0,1)$ aproximadamente. Logo, $\hat{\theta} \mp zK(\hat{\theta})^{1/2}$ corresponde a um intervalo aproximado de $100(1-\alpha)\%$ de confiança para θ, onde z é tal que $\Phi(z) = 1-\alpha/2$, sendo $\Phi(\cdot)$

a função de distribuição acumulada da normal reduzida. A informação observada $J(\hat{\theta})$ poderá substituir $K(\hat{\theta})$ no cálculo deste intervalo. No exemplo 1.2 sobre a distribuição exponencial pode-se calcular diretamente um intervalo de confiança para o parâmetro ρ como $\hat{\rho} \mp z\hat{\rho}/\sqrt{n}$.

1.7 Testes de Hipóteses

A teoria dos testes de hipóteses paramétricos é parte integrante da inferência de verossimilhança e está intimamente relacionada à teoria de estimação. A partir de repetições de um experimento envolvendo um modelo paramétrico, o interesse consiste em determinar se um ou mais parâmetros pertencem a uma dada região do espaço paramétrico. Nos testes paramétricos, as hipóteses são classificadas em *simples* e *compostas*. Se uma distribuição depende de p parâmetros e a hipótese especifica valores para d parâmetros, então ela é *simples* se $d = p$ e *composta* se $d < p$. Em termos geométricos, uma hipótese simples seleciona um único ponto de $\mathbb{R}^d$ enquanto uma hipótese composta corresponde a uma região de $\mathbb{R}^d$ com mais de um ponto. Nas hipóteses compostas, os parâmetros adicionais não-especificados devem ser estimados.

Admite-se que $f(y;\theta)$ é a função de probabilidade conjunta dos dados $y \in \mathbb{R}^n$ e θ é um ponto de $\mathbb{R}^p$. Considere-se uma hipótese nula $H : \theta \in \Theta_0 \subset \Theta$ versus uma alternativa $A : \theta \in \Theta_1 \subset \Theta(\Theta_1 = \Theta - \Theta_0)$. Qualquer teste de hipótese divide o espaço amostral (i.e., o conjunto de valores possíveis do vetor y) em duas regiões mutuamente excludentes: C, a região de rejeição de H (*região crítica*), e $\overline{C}$, a região complementar de aceitação de H. A decisão de um teste consiste em verificar se o vetor de dados y pertence a C ou a $\overline{C}$. Se a distribuição de probabilidade dos dados segundo a hipótese nula H é conhecida, pode-se determinar C tal que, dado H, a probabilidade de rejeitá-la (i.e., $y \in C$) seja menor ou igual a um valor α pré-especificado tal que

$$P(y \in C | \theta \in \Theta_0) \leq \alpha \,. \tag{1.22}$$

A rejeição errônea da hipótese nula H, quando ela é verdadeira, é denominada *erro tipo I*. Assim, a equação (1.22) expressa que a probabilidade do erro tipo I ou alarme falso

nunca excede α (nível de significância do teste). O outro tipo de erro que se pode cometer ao se testar uma hipótese, denominado *erro tipo II*, é função da hipótese alternativa A e representa a aceitação errônea da hipótese nula H quando ela é falsa, sua probabilidade sendo $\beta = P(y \in \overline{C}|\theta \in \Theta_1)$.

Em geral, pode-se encontrar várias regiões críticas satisfazendo (1.22). Qual delas deve ser a preferida? Este é o problema crucial da teoria dos testes de hipóteses. Pode-se escolher uma região crítica C^* tal que ela maximize

$$1 - \beta = P(y \in C|\theta \in \Theta_1) \, .$$

A probabilidade $1 - \beta$, para C fixo, como função do vetor θ especificado na hipótese alternativa, é denominada *função poder* do teste de H versus A.

1.7.1 Hipóteses Simples

Se ambas as hipóteses são simples $\Theta_0 = \{\theta_0\}$ e $\Theta_1 = \{\theta_1\}$, pode-se demonstrar que C^* corresponde ao conjunto de pontos $C^* = \{y; \frac{L(\theta_0)}{L(\theta_1)} \leq k_\alpha\}$, onde k_α é escolhido tal que $\int_C L(\theta_0)dy \leq \alpha$ e $L(\theta)$ é a verossimilhança de θ. A região C^* é considerada a *melhor região crítica* (MRC), pois sua função poder não é menor do que aquela de qualquer outra região satisfazendo (1.22). O teste baseado em C^* é denominado de *teste mais poderoso* (TMP). A razão de verossimilhança $L(\theta_0)/L(\theta_1)$ é uma estatística suficiente quando há apenas duas distribuições em consideração e, portanto, nada mais natural que obter a MRC através desta razão. Quanto menor for esta razão, pior a consistência de H aos dados em questão. Este resultado geral de que a região crítica baseada na razão de verossimilhança produz o TMP de θ_0 versus θ_1 é conhecido como o *Lema de Neyman-Pearson*.

Quando a alternativa a $\theta = \theta_0$ é *unilateral* $\theta_1 > \theta_0$ (ou $\theta_1 < \theta_0$), o mesmo teste também é ótimo para todos os $\theta_1's$ maiores (menores) do que θ_0, sendo denominado de *teste uniformemente mais poderoso* (TUMP). Claramente, esta é uma propriedade mais desejável. Entretanto, quando a alternativa é *bilateral* $\theta_1 \neq \theta_0$ em geral não existe o TUMP. Para obtê-lo, o teste deve estar restrito a certas formas de hipóteses alternativas.

Suponha que existe um vetor S de estatísticas conjuntamente suficientes para um

vetor θ de parâmetros. Comparando-se *duas hipóteses simples* relativas a θ, o teorema da fatoração (1.5) implica $L(\theta_0)/L(\theta_1) = g(s,\theta_0)/g(s,\theta_1)$. Como esperado, se existe a MRC ela é, necessariamente, função dos valores do vetor S segundo H e A. Note-se que a MRC só terá a forma $S \geq a_\alpha$ (ou $S \leq b_\alpha$) quando a razão acima for uma função não-decrescente de s para $\theta_0 > \theta_1$. No caso de θ e s serem escalares, a forma acima ocorrerá quando $\partial^2 \log g(s,\theta)/\partial\theta\partial s \geq 0$. Esta condição é satisfeita para quase todas as distribuições uniparamétricas de probabilidade.

Quando a distribuição dos dados tem mais de um parâmetro e o teste é de *uma hipótese simples* H versus uma *alternativa composta* A, uma MRC variando com os parâmetros segundo A somente existirá em casos especiais. Se existir uma MRC que produza o TUMP de H versus A e um vetor S de estatísticas conjuntamente suficientes para o vetor θ, então a MRC será função de S. Pode-se provar que, se existir um TUMP de H versus A satisfazendo determinadas condições, então existirá um vetor S suficiente para θ. Entretanto, a recíproca em geral não é verdadeira, e a existência de um vetor de estatísticas suficientes não garante a existência de um TUMP para θ.

1.7.2 Hipóteses Compostas

Quando o problema envolve vários parâmetros, a hipótese nula usualmente é composta. Mesmo quando a hipótese nula for simples, a função poder do teste deverá variar com todos os parâmetros, e o ideal seria aumentá-la rapidamente em todas as direções a partir do valor θ_0 especificado na hipótese nula. Entretanto, um sacrifício de declividade, numa dada direção pode aumentar o poder em outra direção. Este dilema só pode ser resolvido ponderando a importância de cada direção de acordo com suas respectivas conseqüências.

Seja $\theta^T = (\psi^T, \lambda^T) \in \mathbb{R}^p$ o vetor de parâmetros particionado em duas componentes. O objetivo é testar a *hipótese nula composta* $H : \psi = \psi^{(0)}$ versus a *hipótese alternativa composta* $A : \psi \neq \psi^{(0)}$, onde ψ e λ são os vetores de interesse e de perturbação, respectivamente, com dimensões q e $p-q$, e $\psi^{(0)}$ é um vetor especificado para ψ. Como a hipótese H não define todas as componentes de θ, o tamanho da região crítica deste teste é função, em geral, dos valores não especificados em λ. Deve-se, então, procurar regiões críticas de tamanhos inferiores a um valor especificado α para todos os valores possíveis

do vetor de perturbação, ou seja, $\alpha(\lambda) \leq \alpha$. No caso de igualdade para todo λ, a região crítica é denominada *similar* para o espaço amostral com respeito a λ. O teste baseado na *região crítica similar* é denominado *teste similar* de tamanho α. Em geral, só existem regiões similares no caso de variáveis aleatórias contínuas *iid*.

Define-se a função característica do conjunto de pontos de uma região C por $\delta(C) = 1$ se $y \in C$ e $\delta(C) = 0$ se $y \notin C$. A esperança matemática $E_Y\{\delta(C)\}$ em relação a Y representa a probabilidade que o ponto amostral y pertença a C e, portanto, é igual ao tamanho de C quando H é verdadeira e a função poder do teste associado a C quando A é verdadeira. Suponha que S é uma estatística suficiente para θ segundo ambas as hipóteses H e A. É fácil mostrar que existe um teste de mesmo tamanho que C baseado em alguma função de S que tem igual poder daquele teste associado à região crítica C. Isto é uma conseqüência imediata do teorema da fatoração (1.5). Note-se que no caso de variáveis contínuas $E_Y\{\delta(C)\} = \int \delta(C)L(\theta)dy$, onde $L(\theta)$ é a verossimilhança de θ. No caso discreto, o somatório substitui a integral. Usando-se (1.5), obtém-se a igualdade, $E_Y\{\delta(C)\} = E_S[E_Y\{\delta(C)|S\}]$, com o operador E_S significando esperança em relação à distribuição de S. Como S é suficiente para θ, $E_Y\{\delta(C)|S\}$ independe de θ e tem a mesma esperança de $\delta(C)$. Logo, existe um teste baseado em S que tem α e β coincidentes com aqueles da região crítica original C. Neste sentido, pode-se restringir, sem perda de poder, a construção dos testes de hipóteses às funções das estatísticas suficientes.

Felizmente, apesar das dificuldades inerentes às hipóteses compostas, existe um método geral para construir regiões críticas em testes de hipóteses compostas, que foi proposto por Neyman e Pearson em 1928. Este método é baseado na razão de verossimilhanças maximizadas segundo ambas hipóteses. No teste de $H : \psi = \psi^{(0)}$ versus $A : \psi \neq \psi^{(0)}$ com o vetor λ desconhecido, seja $L(\psi, \lambda)$ a verossimilhança de ψ e λ. Sejam ainda $\hat{\theta}^T = (\hat{\psi}^T, \hat{\lambda}^T)$ e $\tilde{\theta}^T = (\psi^{(0)^T}, \tilde{\lambda}^T)$ as estimativas de MV de $\theta^T = (\psi^T, \lambda^T)$ correspondentes à maximização de $L(\psi, \lambda)$ segundo A e H, respectivamente. A razão de verossimilhança no teste de H versus A é definida por

$$\ell_R = \frac{L(\psi^{(0)}, \tilde{\lambda})}{L(\hat{\psi}, \hat{\lambda})}, \tag{1.23}$$

e, portanto, representa o quociente entre os máximos das verossimilhanças condicional à

$\psi = \psi^{(0)}$ e incondicional. Evidentemente, $\ell_R \in [0,1]$. Note-se que ℓ_R é uma estatística razoável para testar a hipótese nula H, pois representa a fração do maior valor possível da verossimilhança que é consistente com esta hipótese. Valores grandes de ℓ_R indicam que H é razoável para explicar os dados em questão.

A região crítica do teste é, portanto, $C = \{y; \ell_R \leq k_\alpha\}$, onde k_α é determinado da distribuição (exata ou aproximada) $g(\ell)$ de ℓ_R para produzir um teste de tamanho α, ou seja, $\int_0^{k_\alpha} g(\ell)d\ell = \alpha$. O método da razão de verossimilhança produz regiões críticas similares quando a distribuição de ℓ_R não depende de parâmetros de perturbação. Em geral, isso ocorre num grande número de aplicações. Como a distribuição de ℓ_R é, em geral, complicada, utiliza-se uma transformação conveniente de ℓ_R definida por $w = -2\log \ell_R$ (vide Seção 4.4) que tem, assintoticamente e sob certas condições de regularidade, distribuição χ^2 com graus de liberdade q igual a dimensão do vetor ψ que está sendo testado. A região crítica do teste aproximado de H versus A passa a ser $C = \{y; w \geq \chi^2_q(\alpha)\}$, onde $\chi^2_q(\alpha)$ é o ponto crítico da χ^2_q correspondente ao nível de significância α.

1.8 Exercícios

1. A função de probabilidade de Y em série logarítmica é expressa por $P(Y = y) = \alpha\theta^y/y$ para $0 < \theta < 1$ e $y = 1, 2, \ldots$, onde $\alpha = -\{\log(1-\theta)\}^{-1}$. Demonstre que a EMV de θ é obtida da equação

$$-\hat{\theta}/\{(1-\hat{\theta})\log(1-\hat{\theta})\} = \overline{y},$$

onde $\overline{y}$ é a média amostral.

2. Suponha uma família de densidades indexada por dois parâmetros θ_1 e θ_2. Demonstre que, se t_1 é suficiente para θ_1 quando θ_2 é conhecido e t_2 é suficiente para θ_2 quando θ_1 é conhecido, então (t_1, t_2) é suficiente para (θ_1, θ_2).

3. Suponha a função densidade simétrica em (0,1) dada por $c(\theta)y^\theta(1-y)^\theta$, onde $c(\theta)$ é a inversa da função beta. Calcule a EMV de θ baseada numa amostra de tamanho n. Qual a sua variância assintótica?

4. Obtenha uma estatística t de modo que $P(\sigma^2 \leq t) = 1 - \alpha$ a partir de uma amostra aleatória de tamanho n extraída da distribuição $N(\mu, \sigma^2)$.

5. Considere a função densidade da distribuição gama

$$f(y; \alpha, \phi) = \alpha^{\phi} y^{-1} e^{-\alpha y} / \Gamma(\phi) \ ,$$

onde $\alpha > 0$ e $\phi > 0$. Mostre que as EMV $\hat{\alpha}$ e $\hat{\phi}$ no caso *iid* são calculadas de $\hat{\phi}/\hat{\alpha} = \overline{y}$ e

$$\log \hat{\phi} - \psi(\hat{\phi}) = \log(\overline{y}/\tilde{y}) \ ,$$

onde $\overline{y}$ e $\tilde{y}$ são as médias aritmética e geométrica dos dados e $\psi(x) = d \log \Gamma(x)/dx$ é a função digama.

6. Uma distribuição multinomial tem 4 classes de probabilidades $(1 - \theta)/6$, $(1 + \theta)/6$, $(2 - \theta)/6$ e $(2 + \theta)/6$. Em 1200 ensaios as freqüências observadas nestas classes foram 155, 232, 378 e 435, respectivamente. Calcule a EMV de θ e o seu erro padrão aproximado.

7. Demonstre que a forma mais geral para uma distribuição com parâmetro escalar θ cuja EMV iguala a média aritmética $\overline{y}$ dos dados é $\pi(y; \theta) = \exp\{a(\theta) + a'(\theta)(y - \theta) + c(y)\}$. Assim, $\overline{y}$ é suficiente para θ. Interprete $a(\theta)$. Mostre ainda que se θ é um parâmetro de locação, $\pi(y; \theta)$ é a função densidade da distribuição normal de média θ, e se θ é um parâmetro de escala, $\pi(y; \theta) = \theta^{-1} \exp(-y/\theta)$. Quais seriam as formas da distribuição se no lugar da média aritmética fossem consideradas as médias geométrica e harmônica?

8. Sejam $y_1, \ldots, y_n$ variáveis aleatórias idd com função densidade $\pi(y; \theta)$. Seja $t = t(y_1, \ldots, y_n)$ uma estatística suficiente unidimensional para θ. Se θ_1 e θ_2 são dois valores fixados de θ demonstre que, para todo θ,

$$\frac{\partial}{\partial y} \log \left\{ \frac{\pi(y; \theta)}{\pi(y; \theta_1)} \right\} \Big/ \frac{\partial}{\partial y} \log \left\{ \frac{\pi(y; \theta_2)}{\pi(y; \theta_1)} \right\}$$

é função somente de θ.

9. Sejam $y_1, \ldots, y_n$ uma amostra aleatória de uma distribuição cuja função densidade é

$$f(y;\theta) = (\theta+1)y^{\theta}, \quad y \in (0,1)$$

e $\theta > 0$. (a) Demonstre que a EMV de θ é $\hat{\theta} = -\frac{n}{\sum \log y_i} - 1$; (b) Calcule um intervalo de 95% de confiança para θ.

10. Mostre que as seguintes distribuições são modelos exponenciais da forma (1.18) com $p = 1$ ou $p = 2$: Poisson, binomial, geométrica, gama (índice conhecido), gama (índice desconhecido), Gaussiana inversa e valor extremo. Identifique em cada caso as estatísticas suficientes e os parâmetros canônicos.

11. Sejam $y_1, \ldots, y_n$ observações *iid* de um modelo de locação e escala definido por

$$f(y;\mu,\sigma) = \frac{1}{\sigma} f\left(\frac{y-\mu}{\sigma}\right).$$

(a) Mostre como obter as EMV de μ e σ;

(b) Calcule a matriz de informação para esses parâmetros.

12. A função densidade da distribuição normal inversa com parâmetros $\lambda > 0$ e $\alpha > 0$ é

$$f(y;\alpha,\lambda) = \sqrt{\frac{\lambda}{2\pi}} e^{\sqrt{\lambda\alpha}} y^{-3/2} \exp\left\{-\frac{1}{2}(\lambda y^{-1} + \alpha y)\right\}.$$

(a) Mostre como obter as EMV de α e λ;

(b) Calcule a matriz de informação para esses parâmetros.

Capítulo 2

Métodos Assintóticos

2.1 Conceitos Básicos

O objetivo deste capítulo é apresentar sistematicamente alguns métodos assintóticos úteis em Probabilidade Aplicada e Estatística. O interesse principal é resumir algumas idéias básicas importantes em teoria assintótica e ilustrá-las com aplicações. Os detalhes matemáticos são excluídos e, quando muito, são fornecidas apenas algumas referências e/ou estratégias de demonstração dos resultados. As noções apresentadas neste capítulo formam a base necessária para se entender os demais capítulos deste livro. As seções seguintes exigem que o leitor esteja familiarizado com os conceitos de probabilidade dados aqui. Seja $\{Y_n\}$ uma seqüência de variáveis aleatórias de interesse definida para n grande. Aqui n não representa necessariamente o tamanho da amostra. Apresentam-se inicialmente os quatro modos mais importantes de convergência estocástica.

Convergência em Probabilidade

A seqüência $\{Y_n\}$ converge em probabilidade para uma variável aleatória Y (que pode ser degenerada) se $\lim_{n\to\infty} P(|Y_n - Y| < \epsilon) = 1$ para todo real $\epsilon > 0$. Indica-se esta convergência por $Y_n \xrightarrow{P} Y$. Esta convergência implica, para n suficientemente grande, que Y_n e Y são aproximadamente iguais com probabilidade próxima de 1. O caso especial mais importante é quando $Y_n \xrightarrow{P} k$, onde k é uma constante. Se $h(u)$ é uma função contínua em $u = k$, então $Y_n \xrightarrow{P} k$ implica $h(Y_n) \xrightarrow{P} h(k)$. A noção associada em inferência

estatística é aquela de consistência na estimação de parâmetros.

Se $\{Y_n\}$ é uma seqüência de variáveis aleatórias tal que $E(Y_n) \longrightarrow k$ e $\mathrm{Var}(Y_n) \longrightarrow 0$ quando $n \longrightarrow \infty$, então $Y_n \xrightarrow{P} k$. Entretanto, se $\mathrm{Var}(Y_n) \not\longrightarrow 0$, não se pode tirar qualquer conclusão sobre o comportamento de $\{Y_n\}$. Por exemplo, $E(Y_n) \longrightarrow k$ e $Y_n \xrightarrow{P} k' \neq k$.

Convergência Quase-Certa

Uma seqüência de variáveis aleatórias $\{Y_n\}$ converge quase-certamente (ou converge com probabilidade um) para uma variável aleatória Y se $P\left(\lim_{n\to\infty} Y_n = Y\right) = 1$. Indica-se esta convergência por $Y_n \xrightarrow{q.c.} Y$.

Convergência em Média

Uma seqüência de variáveis aleatórias $\{Y_n\}$ converge em média de ordem r para Y se $\lim_{n\to\infty} E(|Y_n - Y|^r) = 0$. Usa-se a notação $Y_n \xrightarrow{L_r} Y$ para indicar este tipo de convergência. Quanto maior o valor de r mais restritiva é esta condição de convergência. Assim, se $Y_n \xrightarrow{L_r} Y$, então $Y_n \xrightarrow{L_s} Y$ para $0 < s < r$.

Este modo de convergência estocástica admite um critério de convergência. Uma condição necessária e suficiente para $Y_n \xrightarrow{L_r} Y$ é que para todo $\epsilon > 0$ exista um número $n_0 = n_0(\epsilon)$ tal que $|Y_n - Y_m|^r \leq \epsilon$ para quaisquer $m, n \geq n_0$.

As definições de convergência em probabilidade e convergência quase-certa valem para qualquer seqüência de variáveis aleatórias. Entretanto, a convergência em média não vale para qualquer seqüência, pois requer a existência de certos momentos.

Convergência em Distribuição

Uma seqüência de variáveis aleatórias $\{Y_n\}$ converge em distribuição para Y se $\lim_{n\to\infty} P(Y_n \leq y) = F(y)$ para todo ponto y de continuidade da função de distribuição (não-degenerada) F de Y. Para indicar esta convergência usa-se a notação $Y_n \xrightarrow{\mathcal{D}} Y$. Se F é uma função de distribuição degenerada no ponto k, então $P(Y_n \leq y) \longrightarrow 0$ ou

1 dependendo se $y < k$ ou $y \geq k$. Se $h(u)$ é uma função contínua e $Y_n \xrightarrow{\mathcal{D}} Y$, então $h(Y_n) \xrightarrow{\mathcal{D}} h(Y)$.

Dentre as quatro formas de convergência definidas acima, a convergência em distribuição é a mais fraca. Pode-se demonstrar (vide, por exemplo, Wilks, 1962, Capítulo 4, e Serfling, 1980, Capítulo 1) que:

(a) Convergência quase-certa implica convergência em probabilidade;

(b) Convergência em média implica convergência em probabilidade;

(c) Convergência em probabilidade implica convergência em distribuição. As recíprocas das proposições (a) - (c) não são, em geral, verdadeiras;

(d) Se Y é uma variável aleatória degenerada em um ponto k e $Y_n \xrightarrow{\mathcal{D}} Y$, então $Y_n \xrightarrow{P} k$;

(e) Se $\sum_{n=1}^{\infty} P(|Y_n - Y|) > \epsilon) < \infty$ para todo $\epsilon > 0$, então $Y_n \xrightarrow{q.c.} Y$;

(f) Se $\sum_{n=1}^{\infty} E(|Y_n - Y|^r) < \infty$, então $Y_n \xrightarrow{q.c.} Y$;

(g) $Y_n \xrightarrow{\mathcal{D}} Y \not\Rightarrow \lim_{n\to\infty} E(Y_n) = E(Y)$;

(h) $Y_n \xrightarrow{P} Y \not\Rightarrow \lim_{n\to\infty} E(Y_n) = E(Y)$;

(i) Se $Y_n \xrightarrow{L_r} Y$, então $\lim_{n\to\infty} E(|Y_n|^k) = E(|Y|^k)$ para $k \leq r$.

Um caso especial importante de (i) corresponde a $r = 2$. Se $\lim_{n\to\infty} E(|Y_n - Y|^2) = 0$, então $\lim_{n\to\infty} E(Y_n) = E(Y)$ e $\lim_{n\to\infty} E(Y_n^2) = E(Y^2)$;

(j) Se $Y_n \xrightarrow{P} Y$, então existe uma subseqüência $\{Y_{n_j}\}$ de $\{Y_n\}$ tal que $Y_{n_j} \xrightarrow{q.c.} Y$;

(l) $Y_n \xrightarrow{P} Y$ se e somente se toda subseqüência $\{Y_{n_j}\}$ de $\{Y_n\}$ contém uma subseqüência que converge quase certamente para Y.

As convergências em probabilidade e quase-certa não implicam convergência em média. A convergência em distribuição também não implica convergência de momentos e nem mesmo a existência deles. Pode-se comprovar este fato supondo que Y_n tem

função densidade

$$f_n(y) = (1 - e^{-n})\phi(y) + e^{-n}\{\pi(1 + y^2)\}^{-1},$$

onde $\phi(y)$ é a função densidade da normal reduzida. Assim, $f_n(y)$ é uma combinação linear das funções densidades das distribuições normal e Cauchy e converge rapidamente em distribuição para a normal reduzida, mesmo sem seus momentos existirem.

As quatro formas de convergência apresentadas aqui podem ser ilustradas no experimento de infinitos ensaios de Bernoulli independentes. Seja Y_n a proporção de sucessos nas n repetições de Bernoulli independentes, cada uma com probabilidade de sucesso p constante. Tem-se:

$$Y_n \xrightarrow{P} p, \quad Y_n \xrightarrow{q.c.} p,$$

$$\frac{\sqrt{n}(Y_n - p)}{\{p(1-p)\}^{1/2}} \xrightarrow{\mathcal{D}} N(0,1), \quad \frac{\sqrt{n}(Y_n - p)}{(\log\log n)} \xrightarrow{P} 0,$$

$$\frac{\sqrt{n}(Y_n - p)}{(\log\log n)^{1/2}} \not\xrightarrow{q.c.} 0 \quad \text{e} \quad Y_n \xrightarrow{L_2} p.$$

Ordens de Magnitude

Os símbolos $o(\cdot)$ ("de ordem menor que") e $O(\cdot)$ ("de ordem no máximo igual a") são usados para comparar as ordens de magnitude de seqüências de constantes $\{b_n\}, \{c_n\}$. Escreve-se $b_n = o(c_n)$ se $\frac{b_n}{c_n} \longrightarrow 0$ quando $n \longrightarrow \infty$ e $b_n = O(c_n)$ se a razão b_n/c_n é limitada quando $n \longrightarrow \infty$. Assim, supondo n suficientemente grande, $b_n = o(c_n)$ implica que a ordem de magnitude de $\{b_n\}$ é menor que a de $\{c_n\}$, enquanto que $b_n = O(c_n)$ significa que a ordem de magnitude de $\{b_n\}$ é no máximo igual à ordem de $\{c_n\}$. Neste termos, $b_n = o(n^{-1})$ implica que $b_n n \longrightarrow 0$ quando $n \longrightarrow \infty$, enquanto $b_n = O(n^{-1})$ significa que $b_n \leq k/n$ para alguma constante k quando n é suficientemente grande.

As ordens de magnitude acima são trivialmente generalizadas para variáveis aleatórias. Diz-se que $Y_n = o_p(b_n)$ se $\frac{Y_n}{b_n} \xrightarrow{P} 0$. Em especial, $Y_n \xrightarrow{P} k$ é equivalente a $Y_n = k + o_p(1)$. Por outro lado, diz-se que $Y_n = O_p(c_n)$ se a seqüência $\{\frac{Y_n}{c_n}\}$ é limitada em probabilidade para n suficientemente grande. Mais explicitamente, se $Y_n = O_p(c_n)$ então, para todo $\epsilon > 0$, existem constantes k_ϵ e $n_0 = n_0(\epsilon)$ tais que $P(|Y_n| < c_n k_\epsilon) > 1 - \epsilon$ quando $n \geq n_0$. Adicionalmente, se $Y_n \xrightarrow{\mathcal{D}} Y$, então $Y_n = O_p(1)$.

Um caso especial importante é quando $\text{Var}(Y_n) \leq \frac{v}{n}$ se $n > n_0$ para algum $v > 0$ finito. Então, $Y_n = E(Y_n) + O_p(n^{-1/2})$. Se, além disso, $E(Y_n) = \mu + O(n^{-1/2})$ obtém-se o resultado $Y_n = \mu + O_p(n^{-1/2})$, que especifica a taxa de convergência em probabilidade de Y_n para μ.

Mais genericamente, para duas seqüências $\{Y_n\}$ e $\{X_n\}$ de variáveis aleatórias, a notação $Y_n = o_p(X_n)$ significa que $Y_n/X_n \xrightarrow{P} 0$, enquanto $Y_n = O_p(X_n)$ significa que a seqüência $\{Y_n/X_n\}$ é $O_p(1)$.

É fácil verificar que as ordens de magnitude o, O, o_p e O_p satisfazem igualdades tais como: $O(n^{-a})O(n^{-b}) = O(n^{-a-b})$, $O_p(n^{-a})O(n^{-b}) = O_p(n^{-a-b})$, $O_p(n^{-a})o_p(n^{-b}) = o_p(n^{-a-b})$, $o_p(n^{-a})O(n^{-b}) = o_p(n^{-a-b})$, etc.

Normalidade Assintótica

A seqüência $\{Y_n\}$ é assintoticamente normal se existem seqüências de constantes $\{a_n\}, \{b_n\}$ tais que $(Y_n - a_n)/b_n \xrightarrow{\mathcal{D}} Z$, onde Z tem distribuição normal reduzida ($Z \sim N(0,1)$). As constantes a_n, b_n são denominadas média e desvio padrão assintóticos de Y_n, respectivamente. Não há conexão direta entre as constantes a_n, b_n e a média e o desvio padrão de Y_n, embora estas constantes representem realmente em vários casos bem comportados, a média e o desvio padrão de Y_n, respectivamente. Por exemplo, a variável qui-quadrado padronizada $(\chi^2_n - n)/\sqrt{2n}$ é assintoticamente normal. O grande interesse em obter a distribuição normal assintótica é aproximar os quantis da distribuição de Y_n por aqueles da distribuição $N(a_n, b_n^2)$ (vide Seção 3.3).

Embora a normalidade assintótica seja uma característica freqüente e desejável na prática, existem definições similares que se aplicam à convergência para outras distribuições, tais como exponencial, qui-quadrado, Poisson e valor extremo.

Desigualdade de Bienaymé-Chebyshev

Seja Y uma variável aleatória de média e variância finitas. É possível, a partir destes momentos, calcular alguns limites de probabilidade na variabilidade de Y. A desigualdade

de Bienaymé-Chebyshev é expressa (para todo $\epsilon > 0$) como

$$P(|Y - E(Y)| \geq \epsilon \, \mathrm{Var}(Y)^{1/2}) \leq \epsilon^{-2}.$$

Se Y é uma soma de n variáveis aleatórias *iid*, o teorema central do limite (Seção 2.5) mostra que a probabilidade acima tende para $2\Phi(-\epsilon)$ quando $n \longrightarrow \infty$, onde $\Phi(\cdot)$ é a função de distribuição acumulada *(fda)* da distribuição normal $N(0,1)$.

2.2 Função Característica

A função característica de uma variável aleatória Y tendo função de distribuição $F(y)$ é definida por

$$\varphi(t) = E(e^{itY}) = \int_{-\infty}^{+\infty} e^{ity} dF(y), \tag{2.1}$$

onde $i = \sqrt{-1}$ e $t \in \mathbb{R}$. Sejam dois exemplos: para a distribuição de Poisson $P(\lambda)$ de parâmetro λ, $\varphi(t) = \exp\{\lambda(e^{it} - 1)\}$, e para a distribuição normal $N(\mu, \sigma^2)$, $\varphi(t) = \exp(it\,\mu - t^2\sigma^2/2)$.

Supondo certas condições gerais, a função característica determina completamente a função de distribuição. Este fato permite determinar resultados de grande interesse na teoria assintótica. Em inúmeras situações envolvendo funções lineares de variáveis aleatórias independentes, o uso da função característica possibilita determinar a distribuição da função linear em consideração (vide Seção 2.4).

Se o r-ésimo momento μ'_r de Y existe, $\varphi(t)$ pode ser diferenciada k vezes $(0 < k \leq r)$ em relação a t e tem-se

$$\mu'_k = \frac{\varphi^{(k)}(0)}{i^k}, \quad 0 \leq k \leq r,$$

com $\varphi^{(0)}(t) = \varphi(t)$. Assim, $\varphi(t)$ pode ser expandida na vizinhança de $t = 0$ como

$$\varphi(t) = 1 + \sum_{k=1}^{r} \mu'_k \frac{(it)^k}{k!} + o(t^r) \, . \tag{2.2}$$

O logaritmo de $\varphi(t)$ também apresenta uma expansão similar à expansão de $\varphi(t)$

$$\log \varphi(t) = \sum_{k=1}^{r} \kappa_k \frac{(it)^k}{k!} + o(t^r) ,$$

onde os coeficientes $\kappa_k (k = 1, 2, \ldots)$ são denominados de *cumulantes*. Evidentemente, $\kappa_k = \frac{1}{i^k} \frac{d^k \log \varphi(t)}{dt^k}|_{t=0}$ para $0 < k \leq r$. Na Seção 2.3, mostra-se que κ_k é um polinômio em $\mu'_1, \ldots, \mu'_k$ e μ'_k é um polinômio em $\kappa_1, \ldots, \kappa_k$.

Define-se a transformação linear $Z = aY + b$ e sejam $\varphi_Y(t)$ e $\varphi_Z(t)$ as funções características de Y e Z. Mostra-se, facilmente, que

$$\varphi_Z(t) = e^{ibt} \varphi_Y(at) .$$

Em especial, se Z é uma variável aleatória padronizada, isto é, $Z = (Y - \mu)/\sigma$ onde $\mu = E(Y)$ e $\sigma = \mathrm{Var}(Y)^{1/2}$, vem

$$\varphi_Z(t) = \exp\left(-\frac{\mu it}{\sigma}\right) \varphi_Y\left(\frac{t}{\sigma}\right) .$$

Quando $Z = Y + b$, $\varphi_Z(t) = e^{bit}\varphi_Y(t)$ e, então, $\log \varphi_Z(t) = bit + \log \varphi_Y(t)$. Logo, uma translação da variável aleatória Y altera somente o coeficiente de it na expansão de $\log \varphi_Z(t)$, ou seja, os primeiros cumulantes de Z e Y estão relacionados por $\kappa_1(Z) = \kappa_1(Y) + b$, mas os demais cumulantes de Z e Y são iguais $\kappa_r(Z) = \kappa_r(Y)$ para $r \geq 2$. Por causa desta semi-invariância por translação, os cumulantes são também chamados de *semi-invariantes*.

Exemplo 2.1 *Suponha que Y tem distribuição gama ($Y \sim G(p, \alpha)$) com parâmetros p e α, ambos números reais positivos. A função densidade de Y é dada por*

$$f(y) = \alpha^p y^{p-1} e^{-\alpha y}/\Gamma(p) ,$$

onde $\Gamma(p) = \int_0^\infty x^{p-1} e^{-x} dx$ é a função gama definida para x real ou complexo. A função característica segue de

$$\varphi(t) = \frac{\alpha^p}{\Gamma(p)} \int_0^\infty e^{y(-\alpha+it)} y^{p-1} dy .$$

A substituição $z = y(\alpha - it)$ *implica*

$$\varphi(t) = \frac{\alpha^p}{\Gamma(p)(\alpha - it)^p} \int_0^\infty e^{-z} z^{p-1} dz$$

e, finalmente, $\varphi(t) = (1 - \frac{it}{\alpha})^{-p}$. *Assim,*

$$\varphi(t) = 1 + \frac{p}{\alpha} it + \frac{p(p+1)}{\alpha^2} \frac{(it)^2}{2!} + \cdots ,$$

produz os momentos $\mu_1' = p/\alpha$, $\mu_2' = p(p+1)/\alpha^2$, $\mu_3' = p(p+1)(p+2)/\alpha^3$, *etc. Os cumulantes são diretamente obtidos de* $\log \varphi(t)$. *O* k*-ésimo cumulante* κ_k *de* Y *é o coeficiente de* $(it)^k/k!$ *em* $-p\log(1 - \frac{it}{\alpha})$ *e, portanto,* $\kappa_k = (k-1)!p\alpha^{-k}$, $k = 1, 2, \ldots$

Conhecendo a função de distribuição $F(y)$, a função característica pode ser obtida de (2.1). A recíproca também é verdadeira e a função característica determina univocamente a função de distribuição. Em muitos problemas de inferência estatística é mais fácil calcular a função característica do que a correspondente função de distribuição. O problema que surge é como calcular a função de distribuição a partir da função característica. A resposta segue da *fórmula de inversão.*

Assim, dado $\varphi(t)$, a correspondente função de distribuição $F(y)$ é obtida de

$$F(y) - F(0) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \frac{1 - e^{-ity}}{it} \varphi(t) dt , \tag{2.3}$$

suposta contínua em y e 0. Adicionalmente, se $\int_{-\infty}^{+\infty} |\varphi(t)| dt < \infty$, a função característica determina univocamente a função densidade $f(y) = \frac{dF(y)}{dy}$ de uma distribuição contínua por

$$f(y) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} e^{-ity} \varphi(t) dt . \tag{2.4}$$

A demonstração de (2.3) e (2.4) pode ser encontrada em Wilks (1962, p.116), Fisz (1963, p.116) e Rao (1973, p.104). Comparando as fórmulas (2.1) e (2.4) pode ser constatado o tipo de relação recíproca entre $f(y)$ e $\varphi(t)$. Apresentam-se agora dois exemplos de determinação da função densidade a partir da função característica.

Exemplo 2.2 *Obtém-se aqui a função densidade correspondente à função característica*

$\varphi(t) = e^{-t^2/2}$. *Da equação (2.4) vem*

$$\begin{aligned} f(y) &= \frac{1}{2\pi}\int_{-\infty}^{+\infty} e^{-ity}e^{-t^2/2}dt \\ &= \frac{1}{2\pi}\int_{-\infty}^{+\infty} \exp\left\{-\frac{(t+iy)^2}{2}\right\}\exp\left\{\frac{(iy)^2}{2}\right\}dt \\ &= \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{y^2}{2}\right)\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}\exp\left\{-\frac{(t+iy)^2}{2}\right\}dt \end{aligned}$$

e, finalmente, $f(y) = \frac{1}{\sqrt{2\pi}}e^{-y^2/2}$, *que é a função densidade da distribuição normal reduzida.*

Exemplo 2.3 *Deseja-se calcular a função densidade correspondente à função característica* $\varphi(t) = e^{-|t|}$. *De (2.4) vem*

$$f(y) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} e^{-ity}e^{-|t|}dy$$

e, por simetria,

$$\pi f(y) = \int_0^{\infty} e^{-t}\cos(ty)\,dt = -e^{-t}\cos(ty)\Big|_0^{\infty} - y\int_0^{\infty} e^{-t}\mathrm{sen}(ty)\,dt = 1 - y^2\pi f(y)\,.$$

Logo, $f(y) = \frac{1}{\pi(1+y^2)}$, $y \in \mathbb{R}$, *que é a função densidade da distribuição de Cauchy.*

A equação (2.3) contém $F(0)$ e a determinação desta quantidade pode ser evitada usando a fórmula de inversão alternativa

$$F(y) = \frac{1}{2} + \frac{1}{2\pi}\int_0^{\infty}\{e^{ity}\varphi(-t) - e^{-ity}\varphi(t)\}\frac{dt}{it}\,.$$

No caso de distribuições discretas nos inteiros não negativos, a fórmula correspondente à equação (2.4) é

$$P(Y = y) = \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{-ity}\varphi(t)dt,$$

com alteração apenas nos limites de integração.

Como a função característica determina univocamente a função de distribuição, o problema de convergência em probabilidade de uma seqüência de variáveis aleatórias

pode ser resolvido através da convergência da seqüência correspondente de funções características. Este princípio fundamental, de grande interesse na teoria assintótica, é conhecido como o *teorema da continuidade* (Levy, 1937; Cramér, 1937), descrito abaixo.

Teorema da Continuidade

Seja $\{Y_n\}$ uma seqüência de variáveis aleatórias tendo funções de distribuição $F_1, F_2, \ldots$ e com funções características correspondentes $\varphi_1, \varphi_2, \ldots$ Se φ_n converge pontualmente para um limite φ e se φ é contínua no ponto zero, então existe uma função de distribuição F de uma variável aleatória Y tal que $Y_n \xrightarrow{\mathcal{D}} Y$, sendo φ a função característica de Y.

Da definição de convergência em distribuição de uma seqüência $\{Y_n\}$ de variáveis aleatórias, i.e., $Y_n \xrightarrow{\mathcal{D}} Y$, usa-se também uma notação equivalente $F_n \longrightarrow F$ para as funções de distribuição de Y_n e Y.

Corolário

Supondo que as funções de distribuição $F, F_1, F_2, \ldots$ têm funções características correspondentes $\varphi, \varphi_1, \varphi_2, \ldots$, então as seguintes proposições são equivalentes:

i) $F_n \longrightarrow F$;

ii) $\lim_{n\to\infty} \varphi_n(t) = \varphi(t)$, para qualquer $t \in \mathbb{R}$, e $\varphi(t)$ sendo contínua em $t = 0$;

iii) $\lim_{n\to\infty} \int g dF_n = \int g dF$, sendo g uma função contínua limitada, i.e., $|g| < c$ para algum $c \in \mathbb{R}$.

Se $F_n \longrightarrow F$, e F é contínua, então a convergência é uniforme, ou seja, $\lim_{n\to\infty} \sup_x |F_n(x) - F| = 0$.

2.3 Momentos e Cumulantes

As funções geratrizes de momentos (fgm) e de cumulantes (fgc) de Y são definidas por $M(t) = E(e^{tY})$ e $K(t) = \log M(t)$, respectivamente. Observe-se que a função característica $\varphi(t)$ é expressa diretamente pela fgm $M(t)$ através de $\varphi(t) = M(it)$. Quando a fgm não converge para t real num intervalo contendo a origem, trabalha-se geralmente

com a função característica, que existe sempre para t real e determina univocamente a distribuição. Evidentemente, $M(t)$ e $K(t)$ têm a mesma propriedade geradora de momentos e cumulantes que $\varphi(t)$ e $\log \varphi(t)$, respectivamente. Com efeito, $\mu'_r = M^{(r)}(0)$ e $\kappa_r = K^{(r)}(0)$, onde o sobrescrito (r) indica a r-ésima derivada em relação a t.

Exemplo 2.4 *Para a distribuição normal $N(\mu, \sigma^2)$ obtém-se, facilmente,*

$$M(t) = \exp\left(t\mu + \frac{1}{2}t^2\sigma^2\right)$$

e, então, $K(t) = \mu t + \frac{1}{2}\sigma^2 t^2$, de modo que $\kappa_1 = \mu$, $\kappa_2 = \sigma^2$ e $\kappa_r = 0$ para $r \geq 3$. Como todos os cumulantes da normal, acima de segunda ordem, são nulos, a proximidade de uma distribuição pela distribuição normal pode ser determinada pelas magnitudes de seus cumulantes. Este fato revela a importância dos cumulantes na teoria assintótica.

Exemplo 2.5 *Suponha que Y tem função densidade na família exponencial*

$$f(y) = \exp\{y\theta - b(\theta) + a(y)\},\ y \in I\!R_Y\ . \tag{2.5}$$

A condição de normalização

$$\int_{I\!R_Y} \exp\{y\theta - b(\theta) + a(y)\}dy = 1$$

implica para todo θ

$$M(t) = \int \exp\{yt + y\theta - b(\theta) + a(y)\}dy$$

e, então, a fgm de Y é dada por

$$M(t) = \exp\{b(\theta + t) - b(\theta)\}\ .$$

A fgc de Y segue como $K(t) = \log M(t) = b(\theta+t) - b(\theta)$. Logo, o r-ésimo cumulante de Y é dado por $\kappa_r = K^{(r)}(0) = b^{(r)}(\theta)$. Assim, a função $b(\theta)$ na família exponencial (2.5) gera os cumulantes de Y. A função $b(\theta)$ está relacionada diretamente à log-verossimilhança de θ e este fato representa uma das maiores motivações para o uso de cumulantes na teoria assintótica.

Seja Y uma variável aleatória e $Z = aY + b$ uma transformação linear de Y. É fácil verificar que os r-ésimos cumulantes de $Z\,(\kappa_r(Z))$ e $Y\,(\kappa_r(Y))$ são expressos por $\kappa_r(Z) = a^r\kappa_r(Y)$. Assim, os *cumulantes padronizados* de Z e Y definidos por $\rho_r = \kappa_r/\kappa_2^{r/2}$ são iguais, i.e., $\rho_r(Z) = \rho_r(Y)$. Logo, os cumulantes padronizados de variáveis aleatórias são invariantes segundo transformações lineares. Os momentos têm uma vantagem sobre os cumulantes devido à interpretação (física e geométrica) simples. Entretanto, os cumulantes na teoria assintótica são de maior interesse que os momentos, principalmente porque se anulam para a distribuição normal e, com uma simples padronização, se tornam invariantes segundo transformações lineares. Mostra-se, a seguir, que o conhecimento de momentos e de cumulantes até uma dada ordem são equivalentes.

A função geratriz de momentos $M(t)$ pode ser representada pela expansão

$$M(t) = 1 + \sum_k \mu'_k \frac{t^k}{k!}\,, \tag{2.6}$$

suposta convergente para todo $|t|$ suficientemente pequeno. A soma ilimitada em (2.6) pode ser divergente para todo real $|t| > 0$ porque alguns dos momentos de ordem superior são infinitos ou porque os momentos, embora finitos, aumentam rapidamente, forçando a divergência. Neste caso, trabalha-se com expansões finitas até um certo número de termos, especificando a ordem do erro como função do tamanho da amostra n ou de alguma quantidade relacionada a n.

A função geratriz de cumulantes é expandida como

$$K(t) = \sum_k \kappa_k \frac{t^k}{k!}\,. \tag{2.7}$$

Das equações (2.6) e (2.7) vem

$$\exp\left(\sum_k \kappa_k \frac{t^k}{k!}\right) = 1 + \sum_k \mu'_k \frac{t^k}{k!}\,.$$

Expandindo em série de Taylor a função exponencial anterior e igualando os coeficientes de mesma potência em t, expressam-se os momentos em termos dos cumulantes

de mesma ordem e de ordem inferior. Os seis primeiros momentos são:

$$\mu_1' = \kappa_1,\ \mu_2' = \kappa_2 + \kappa_1^2,\ \mu_3' = \kappa_3 + 3\kappa_2\kappa_1 + \kappa_1^3,\ \mu_4' = \kappa_4 + 4\kappa_3\kappa_1 + 3\kappa_2^2 + 6\kappa_2\kappa_1^2 + \kappa_1^4,$$
$$\mu_5' = \kappa_5 + 5\kappa_4\kappa_1 + 10\kappa_3\kappa_2 + 10\kappa_3\kappa_1^2 + 15\kappa_2^2\kappa_1 + 10\kappa_2\kappa_1^3 + \kappa_1^5,\ \mu_6' = \kappa_6 + 6\kappa_5\kappa_1 + 15\kappa_4\kappa_2$$
$$+15\kappa_4\kappa_1^2 + 10\kappa_3^2 + 60\kappa_3\kappa_2\kappa_1 + 20\kappa_3\kappa_1^3 + 15\kappa_2^3 + 45\kappa_2^2\kappa_1^2 + 15\kappa_2\kappa_1^4 + \kappa_1^6.$$

A inversão das equações acima pode ser obtida diretamente destas fórmulas ou, mais facilmente, expandindo o logaritmo abaixo em série de Taylor

$$\sum_k \kappa_k \frac{t^k}{k!} = \log\left\{1 + \sum_k \mu_k' \frac{t^k}{k!}\right\}$$

e igualando os coeficientes de mesma potência em t. Encontram-se,

$$\kappa_1 = \mu_1',\ \kappa_2 = \mu_2' - \mu_1'^2,\ \kappa_3 = \mu_3' - 3\mu_2'\mu_1' + 2\mu_1'^3,\ \kappa_4 = \mu_4' - 4\mu_3'\mu_1' - 3\mu_2'^2 + 12\mu_2'\mu_1'^2$$
$$-6\mu_1'^4,\ \kappa_5 = \mu_5' - 5\mu_4'\mu_1' - 10\mu_3'\mu_2' + 20\mu_3'\mu_1'^2 + 30\mu_2'^2\mu_1' - 60\mu_2'\mu_1'^3 + 24\mu_1'^5,$$
$$\kappa_6 = \mu_6' - 6\mu_5'\mu_1' - 15\mu_4'\mu_2' + 30\mu_4'\mu_1'^2 - 10\mu_3'^2 + 120\mu_3'\mu_2'\mu_1' - 120\mu_3'\mu_1'^3 + 30\mu_2'^3$$
$$-270\mu_2'^2\mu_1'^2 + 360\mu_2'\mu_1'^4 - 120\mu_1'^6.$$

Assim, existe uma relação biunívoca entre momentos e cumulantes. Entretanto, os cumulantes oferecem mais vantagens em termos estatísticos do que os momentos. Entre estas vantagens, citam-se: (a) muitos cálculos estatísticos usando cumulantes são mais fáceis do que os cálculos correspondentes através de momentos; (b) para variáveis aleatórias independentes, os cumulantes de uma soma são, simplesmente, somas dos cumulantes das variáveis individuais; (c) séries do tipo Edgeworth para aproximar densidades e distribuições (vide Seção 3.3) e logaritmos de densidades são expressas de forma mais conveniente via cumulantes ao invés de momentos; (d) os cumulantes de inúmeras distribuições podem ter ordens pré-estabelecidas, o que não ocorre com os momentos; (e) considerando a aproximação normal (vide Seções 3.3 e 3.10), os cumulantes (mas não os momentos) de ordem superior a um valor especificado podem, usualmente, ser ignorados, pois tendem a zero mais rapidamente que os demais quando o tamanho da amostra cresce.

Além destas vantagens, os cumulantes têm interpretação simples. Os dois primeiros cumulantes são o valor médio e a variância da variável Y em consideração. O tercei-

ro cumulante é uma medida de assimetria da distribuição de Y no sentido de que κ_3 é zero quando Y é distribuída simetricamente. Entretanto, $\kappa_3 = 0$ não é uma condição suficiente para Y ter distribuição simétrica. Para termos simetria a distribuição deve ser univocamente determinada pelos seus cumulantes e todos os cumulantes de ordem ímpar devem se anular. O quarto cumulante é uma medida de curtose da distribuição de Y. Os cumulantes de ordem superior a quatro podem ser interpretados como medidas de não-normalidade, pois eles se anulam quando Y tem distribuição normal. Na teoria assintótica os cumulantes padronizados $\rho_r = \kappa_r/\kappa_2^{r/2}$, para $r = 1, 2, \ldots$, são mais importantes, principalmente ρ_3 e ρ_4, por causa da invariância segundo transformação linear e por terem ordens pré-estabelecidas.

Em muitas situações é mais fácil trabalhar com momentos centrais (μ_r) do que com momentos ordinários (μ'_r). Existem relações simples de recorrência entre esses momentos. Tem-se $\mu_r = E\{(Y - \mu'_1)^r\}$ e desenvolvendo o binômio vem:

$$\mu_r = \sum_{k=0}^{r} \binom{r}{k} \mu'_{r-k}(-\mu'_1)^k \ .$$

Analogamente,

$$\mu'_r = \sum_{k=0}^{r} \binom{r}{k} \mu_{r-k}\mu'^{k}_1 \ .$$

Em especial, relações entre cumulantes e momentos centrais são bem mais simples do que entre cumulantes e momentos ordinários. As seis primeiras são:

$$\mu_1 = 0, \ \mu_2 = \kappa_2, \ \mu_3 = \kappa_3, \ \mu_4 = \kappa_4 + 3\kappa_2^2, \ \mu_5 = \kappa_5 + 10\kappa_2\kappa_3,$$
$$\mu_6 = \kappa_6 + 15\kappa_2\kappa_4 + 10\kappa_3^2 + 15\kappa_2^3$$

e

$$\kappa_2 = \mu_2, \ \kappa_3 = \mu_3, \ \kappa_4 = \mu_4 - 3\mu_2^2, \ \kappa_5 = \mu_5 - 10\mu_2\mu_3,$$
$$\kappa_6 = \mu_6 - 15\mu_2\mu_4 - 10\mu_3^2 + 30\mu_2^3.$$

Exemplo 2.6 *Suponha que Y tem distribuição binomial $B(n,p)$ com parâmetros n e p. Tem-se $M(t) = (1-p+pe^t)^n$, $K(t) = n\log(1-p+pe^t)$ e $\varphi(t) = M(it) = (1-p+pe^{it})^n$. Calculam-se, facilmente, $\kappa_1 = \mu'_1 = np$, $\kappa_2 = \mu_2 = np(1-p)$, $\kappa_3 = \mu_3 = np(1-p)(1-2p)$, $\mu_4 = 3n^2p^2(1-p)^2 + np(1-p)(1-6p+6p^2)$ e $\kappa_4 = np(1-p)(1-6p+6p^2)$.*

Assim, os cumulantes padronizados $\rho_3 = \kappa_3/k_2^{3/2}$ *e* $\rho_4 = \kappa_4/\kappa_2^2$ *de* Y *são*

$$\rho_3 = \frac{1-2p}{\sqrt{np(1-p)}} \quad e \quad \rho_4 = \frac{1-6p+6p^2}{np(1-p)} .$$

Note-se que ρ_3 *e* $\rho_4 \longrightarrow 0$ *quando* $n \longrightarrow \infty$. *Este resultado está de acordo com o teorema de DeMoivre-Laplace (Seção 2.5) que mostra que a distribuição binomial padronizada tende para a distribuição normal quando* $n \longrightarrow \infty$.

Na Seção 2.2 mostrou-se que os momentos de uma variável aleatória, se existirem, podem ser calculados derivando a função característica e que, também, a função característica determina a distribuição. Entretanto, isto não implica que o conhecimento dos momentos determine completamente a distribuição, mesmo quando os momentos de todas as ordens existem. Somente segundo certas condições, que felizmente são satisfeitas para as distribuições comumente usadas na teoria assintótica, é que um conjunto de momentos determina univocamente a distribuição. Em termos práticos, o conhecimento de momentos, quando todos eles existem, é em geral equivalente ao conhecimento da distribuição, no sentido de que é possível expressar todas as propriedades da distribuição em termos de momentos.

Em algumas situações, os momentos são mais facilmente obtidos através de outros métodos que não o de derivar $\varphi(t)$ ou $M(t)$. Uma pergunta pertinente é: Quais as condições para que uma seqüência de momentos $\mu'_1, \mu'_2, \ldots$ de uma variável aleatória Y determine univocamente a função de distribuição de Y? Uma condição suficiente devida a Cramér (1946) é a seguinte. Seja $F(y)$ uma função de distribuição cujos momentos μ'_k, $k = 1, 2, \ldots$, são todos finitos. Se a série $\sum_{k=0}^{\infty} \frac{\mu'_k t^k}{k!}$ é absolutamente convergente para algum $t > 0$, então $F(y)$ é a única função de distribuição cujos momentos correspondentes são iguais a μ'_k, $k = 1, 2, \ldots$

No caso da variável aleatória ser limitada, i.e., se existirem números a e b finitos $(a < b)$ tais que $F(a) = 0$ e $F(b) = 1$, então sua função de distribuição $F(y)$ é univocamente determinada pelos momentos μ'_k, $k = 1, 2, \ldots$, desde que todos eles existam.

Uma dificuldade que surge no cálculo de momentos e cumulantes para demonstrar resultados de natureza genérica em teoria assintótica é que o conjunto infinito de momentos

(ou cumulantes), pode não ser suficiente para determinar a distribuição univocamente. Por exemplo, Feller (1971, Seção VII.3) apresenta um par de funções densidades distintas produzindo momentos idênticos de todas as ordens. A não-unicidade ocorre quando a função $M(t)$ não é analítica na origem. Em um grande número de problemas, pode-se evitar a não-unicidade incluindo a condição de que a expansão (2.6) seja convergente para $|t| < \delta$, onde $\delta > 0$.

Finalmente, suponha que $\{Y_n\}$ é uma seqüência de variáveis aleatórias com funções de distribuição $F_1, F_2, \ldots$ e cujas seqüências de momentos são conhecidas. Seja μ'_{rn} o r-ésimo momento de Y_n, suposto finito para quaisquer n e r. Apresenta-se, agora, um critério simples baseado em momentos para determinar se $Y_n \xrightarrow{\mathcal{D}} Y$. Suponha que $\lim_{n\to\infty} \mu'_{r,n} = \mu'_r$, onde μ'_r é finito para todo r. Se $F_n \longrightarrow F$, então $\mu'_0, \mu'_1, \mu'_2, \ldots$ é a seqüência de momentos correspondente a F. Em sentido inverso, se $\mu'_0, \mu'_1, \mu'_2, \ldots$ determina univocamente a distribuição $F(y)$, então $F_n \longrightarrow F$. A demonstração deste resultado pode ser encontrada em Kendall e Rao (1950).

2.4 Somas de Variáveis Aleatórias Independentes

O cálculo de distribuições assintóticas de somas de variáveis aleatórias independentes é muito freqüente em inferência estatística. Esta seção trata de algumas propriedades das somas de variáveis aleatórias independentes supondo um número n finito dessas variáveis. Na Seção 2.5 e no Capítulo 3 consideram-se propriedades das somas quando $n \longrightarrow \infty$.

Sejam $Y_1, \ldots, Y_n$ variáveis aleatórias *iid*, cópias de uma variável aleatória Y. Seja $S_n = \sum_{i=1}^{n} Y_i$ a soma das n variáveis supondo que todos os momentos de Y existem e que $E(Y) = \mu$ e $\mathrm{Var}(Y) = \sigma^2$. Tem-se $E(S_n) = n\mu$ e $\mathrm{Var}(S_n) = n\sigma^2$.

Em cálculos estatísticos é comum padronizar a variável aleatória de interesse de modo que uma distribuição limite não-degenerada seja obtida quando $n \longrightarrow \infty$. Em geral, padroniza-se a nova variável de modo que ela tenha, exatamente ou aproximadamente, média zero e variância constante, ou mesmo unitária. Assim, obtém-se a soma padronizada $S_n^* = (S_n - n\mu)/(\sqrt{n}\sigma)$, que satisfaz $E(S_n^*) = 0$ e $\mathrm{Var}(S_n^*) = 1$.

A *fgm* $M_{S_n}(t)$ de S_n é calculada a partir da *fgm* $M_Y(t)$ de Y através de

$$M_{S_n}(t) = E(e^{tS_n}) = \prod_{i=1}^{n} E(e^{tY_i}) = M_Y(t)^n$$

e, portanto, a *fgc* $K_{S_n}(t)$ é simplesmente um múltiplo da *fgc* $K_Y(t)$

$$K_{S_n}(t) = n\, K_Y(t)\ . \tag{2.8}$$

Logo, os cumulantes de S_n são simplesmente iguais a n vezes os respectivos cumulantes de Y, ou seja,

$$\kappa_r(S_n) = n\,\kappa_r(Y) \tag{2.9}$$

para $r \geq 1$. A equação (2.9) apresenta um forte motivo para se trabalhar com cumulantes no contexto de somas de variáveis aleatórias *iid*. Da equação (2.9) obtêm-se os cumulantes padronizados de S_n como

$$\rho_3(S_n) = \frac{\rho_3(Y)}{\sqrt{n}},\ \rho_4(S_n) = \frac{\rho_4(Y)}{n},\ \rho_r(S_n) = \frac{\rho_r(Y)}{n^{r/2-1}}$$

e, assim, estes cumulantes decrescem em potências de $1/\sqrt{n}$. Este fato também é muito importante no desenvolvimento das expansões de Edgeworth apresentadas na Seção 3.3. Os cumulantes padronizados de S_n^* são iguais aos correspondentes cumulantes de S_n devido à invariância segundo uma transformação linear.

A função densidade exata de S_n (ou S_n^*) pode ser calculada pela *convolução* (soma ou integral), quando n é pequeno. Assim, no caso contínuo, onde as variáveis são *iid* com função densidade $f_Y(y)$, a função densidade $f_{S_n}(s)$ de S_n é expressa pela integral múltipla de dimensão $n-1$

$$f_{S_n}(s) = \int \left\{ \prod_{i=1}^{n-1} f_{Y_i}(y_i) \right\} f_{Y_n}\left(s - \sum_{i=1}^{n-1} y_i \right) \prod_{i=1}^{n-1} dy_i\ .$$

No caso discreto esta integral deve ser substituída por um somatório. As funções de distribuição de S_n e S_n^* seguem de $F_{S_n}(z) = \int_{-\infty}^{z} f_{S_n}(s)ds$ e $F_{S_n^*}(z) = F_{S_n}(n\mu + \sqrt{n}\sigma z)$, respectivamente.

O cálculo algébrico da função densidade de S_n pela fórmula da convolução só é útil para valores pequenos de n ou em casos especiais. Na prática é mais comum determinar a distribuição exata de S_n a partir da fórmula de inversão (Seção 2.2) ou do critério de reprodutividade da função característica dado a seguir, ou então através das aproximações assintóticas quando $n \longrightarrow \infty$ (vide Seção 2.5 e Capítulo 3).

Para a determinação numérica da integral relativa à $f_{S_n}(s)$ dada anteriormente aproxima-se, em geral, a função densidade $f_Y(y)$ de Y por uma função densidade conhecida $g(y)$, onde as convoluções podem ser calculadas explicitamente em forma simples. Considera-se, assim, $f_Y(y) = g(y) + \delta(y)$, onde $\delta(y)$ é uma pequena perturbação. Em especial, a escolha de $\delta(y)$ pode ser $\delta(y) = g(y)\sum_r c_r p_r(y)$, onde $\{p_r(y)\}$ é um conjunto de polinômios ortogonais associados a $g(y)$ (vide Seção 3.2). Neste caso, pode-se ter uma expansão para a convolução onde os termos principais são facilmente calculados.

No contexto das aplicações, as funções características fornecem os métodos mais poderosos para determinar a função de distribuição de somas (e médias) de variáveis aleatórias independentes. Em especial, a função característica $\varphi_{S_n}(t)$ de S_n tem a propriedade do produto linear similar àquela de $M_{S_n}(t)$. Assim, no caso de variáveis aleatórias independentes $Y_1, \ldots, Y_n$ com funções características respectivas $\varphi_1(t), \ldots, \varphi_n(t)$, a função característica de $S_n = \sum_{i=1}^{n} Y_i$ é dada por

$$\varphi_{S_n}(t) = \prod_{i=1}^{n} \varphi_i(t) \,. \tag{2.10}$$

Quando as variáveis aleatórias são *iid*, as funções características de S_n e da média $\overline{Y}_n = S_n/n$ são iguais a $\varphi(t)^n$ e $\varphi(\frac{t}{n})^n$, respectivamente, e a função característica de S_n^* segue de

$$\varphi_{S_n^*}(t) = \exp\left(-\frac{\sqrt{n}\mu it}{\sigma}\right)\varphi\left(\frac{t}{\sqrt{n}\sigma}\right)^n \,. \tag{2.11}$$

O resultado (2.10) da função característica de uma soma de variáveis aleatórias independentes é facilmente estendido para uma combinação linear $Z = \sum_{i=1}^{k} c_i Y_i$. Sendo $\varphi_i(t_i)$ a função característica de Y_i, $i = 1, \ldots, k$, tem-se $\varphi_Z(t) = \prod_{i=1}^{k} \varphi_i(c_i t)$.

A função de distribuição de S_n (ou $\overline{Y}_n$), pelo menos em teoria, pode ser determinada a partir da sua função característica em (2.10) usando a integral (2.3), embora em certos casos a avaliação desta integral seja difícil. Em muitas situações onde as variáveis aleatórias são *iid*, a determinação das funções de distribuição de S_n e $\overline{Y}_n$ pode ser feita a partir do *critério de reprodutividade* da função característica. Segundo este critério, se $\varphi_Y(t;\theta)$ é a função característica de Y, que depende de um certo vetor θ de parâmetros da sua distribuição, então a função característica de S_n pode ser expressa por

$$\varphi_{S_n}(t;\theta) = \varphi_Y(t;\theta)^n = \varphi_Y(t;n\theta) \ .$$

No caso do critério acima ser satisfeito, S_n tem a mesma distribuição de Y a menos da permuta do vetor de parâmetros θ por $n\theta$. Por exemplo, baseando-se neste critério, é fácil mostrar que se Y tem distribuição $B(m,p)$, $P(\mu)$ e $N(\mu,\sigma^2)$, então $S_n = \sum_{i=1}^{n} Y_i$ tem distribuição $B(nm,p)$, $P(n\mu)$ e $N(n\mu, n\sigma^2)$, respectivamente.

2.5 Teoremas Limites

A Seção 2.4 tratou do cálculo da distribuição de uma soma de variáveis aleatórias *iid* supondo n fixo. Esta seção apresenta resultados importantes sobre a distribuição da soma de variáveis aleatórias *iid* quando $n \longrightarrow \infty$. Estes resultados consistem em teoremas limites bastante úteis na inferência para aproximar distribuições de estatísticas (em grandes amostras) pela distribuição normal. Nas aplicações verifica-se que muitos desses resultados assintóticos fornecem boas aproximações em amostras moderadas. Os teoremas limites mais citados são aqueles de Lindeberg-Lévy, Liapunov, Lindeberg-Feller e a integral de DeMoivre-Laplace. A grande maioria destes teoremas foi desenvolvida entre 1920 e 1945 por B.W. Gnedenko, A. Khintchin, P. Lévy, J.W. Lindeberg e A.N. Kolmogorov. Um estudo detalhado pode ser encontrado em Wilks (1962, Capítulo 9), Fisz (1963, Capítulo 6), Feller (1971, Capítulo VIII) e Rao (1973, Seção 2c).

Seja $\{Y_n\}$ uma seqüência de variáveis aleatórias *iid*, $S_n = \sum_{n=1}^{n} Y_i$ a soma das n primeiras observações e $\overline{Y}_n = S_n/n$ a sua média. Quando se conhece apenas a média $E(Y_i) = \mu$

da seqüência, as conclusões sobre o comportamento de $\overline{Y}_n$ para n grande são dadas pelas *Leis Fraca* e *Forte dos Grandes Números apresentadas a seguir:*

Lei Fraca dos Grandes Números
Se existe $E(Y_i) = \mu < \infty$, então $\overline{Y}_n \xrightarrow{P} \mu$.

Lei Forte dos Grandes Números
Uma condição necessária e suficiente para $\overline{Y}_n \xrightarrow{q.c.} \mu$ é que exista $E(Y_i)$ e $E(Y_i) = \mu$.

Quando se conhece a média $E(Y_i) = \mu$ e a variância $\text{Var}(Y_i) = \sigma^2$ da seqüência, pode-se trabalhar com o teorema central do limite, que mostra o papel de destaque da distribuição normal na teoria assintótica. O teorema central do limite é um nome genérico para qualquer teorema dando a convergência (em distribuição) de uma soma de variáveis aleatórias para a distribuição normal. Formas clássicas deste teorema se referem à soma de variáveis aleatórias independentes. No contexto de teoremas limites algumas vezes usam-se os termos "global" e "local" para se referir às convergências das funções de distribuição e densidade, respectivamente. O termo "teorema limite local" é também usado quando uma função de probabilidade discreta é aproximada por uma função densidade (vide teorema de DeMoivre-Laplace a seguir). Se, além da média μ, a variância σ^2 da seqüência $\{Y_n\}$ é também conhecida, pode-se obter mais informação sobre o comportamento de $\overline{Y}_n$ quando $n \longrightarrow \infty$. No contexto de variáveis aleatórias *iid*, o teorema central do limite de Lindeberg-Lévy representa a forma mais simples dos teoremas centrais de limite mais gerais.

Teorema de Lindeberg-Lévy

Seja $S_n^* = \frac{S_n - n\mu}{\sqrt{n}\sigma} = \frac{\sqrt{n}}{\sigma}(\overline{Y}_n - \mu)$ a soma padronizada de n variáveis aleatórias *iid*. Se os dois primeiros momentos $E(Y_i) = \mu$ e $\text{Var}(Y_i) = \sigma^2$ existem e ambos são finitos, então $S_n^* \xrightarrow{D} N(0,1)$, i.e.,

$$\lim_{n\to\infty} P(S_n^* \leq y) = \Phi(y) \; . \tag{2.12}$$

Como a distribuição limite é contínua, a convergência da função de distribuição de S_n^* para $\Phi(\cdot)$ é uniforme e, então,

$$\{P(S_n^* \leq t_n) - \Phi(t_n)\} \longrightarrow 0$$

quando $n \longrightarrow \infty$, onde t_n pode depender de n de qualquer forma.

Seja $\rho(t)$ a função característica de $Y_i - \mu$. Como $E(Y_i - \mu) = 0$ e $\mathrm{Var}(Y_i - \mu) = \sigma^2$ obtém-se, expandindo $\rho(t)$ como em (2.2),

$$\rho(t) = 1 - \frac{\sigma^2 t^2}{2} + o(t^2) \ .$$

A função característica de $S_n^* = \sum_{i=1}^{n}(Y_i - \mu)/\sigma\sqrt{n}$ é $\varphi_{S_n^*}(t) = \rho(\frac{t}{\sigma\sqrt{n}})^n$. Logo,

$$\varphi_{S_n^*}(t) = \left\{1 - \frac{t^2}{2n} + o\left(\frac{t^2}{n}\right)\right\}^n$$

e, portanto, $\lim_{n\to\infty} \varphi_{S_n^*}(t) = e^{-t^2/2}$. Como $e^{-t^2/2}$ é a função característica da distribuição normal $N(0,1)$, a equação (2.12) decorre do teorema da continuidade.

A convergência $S_n^* \xrightarrow{\mathcal{D}} N(0,1)$, ou equivalentemente, $\sigma^{-1}\sqrt{n}(\overline{Y}_n - \mu) \xrightarrow{\mathcal{D}} N(0,1)$ representa o resultado central da teoria estatística, pois permite construir intervalos de confiança aproximados e testar hipóteses sobre μ usando a média amostral $\overline{Y}_n$ e sua distribuição normal $N(\mu, \frac{\sigma^2}{n})$ aproximada.

A equação (2.12) garante que a *fda* da soma padronizada S_n^* converge para a distribuição normal reduzida. Entretanto, a função densidade de S_n^* não converge necessariamente para a função densidade da distribuição normal reduzida, pois as variáveis aleatórias $Y_1, \ldots, Y_n$ podem ser discretas. Há condições bem gerais que garantem que a função de probabilidade de S_n^* pode ser aproximada no caso discreto pela função densidade $\phi(x) = \frac{1}{\sqrt{2\pi}}e^{-x^2/2}$ da distribuição $N(0,1)$. O leitor deve consultar na Seção 3.10 as aproximações baseadas na distribuição normal para algumas variáveis aleatórias discretas.

Teorema Central do Limite (Local) para Densidades

Seja $F_{S_n^*}(y) = P(S_n^* \leq y)$ a *fda* de S_n^* no contexto de variáveis aleatórias *iid.* Então, S_n^* tem uma função densidade contínua $f_{S_n^*}(y) = \frac{dF_{S_n^*}(y)}{dy}$ para todo n suficientemente grande e

$$\lim_{n\to\infty} f_{S_n^*}(y) = \phi(y) \tag{2.13}$$

uniformemente em $y \in \mathbb{R}$ se, e somente se, existir um inteiro $k > 0$ para o qual a função característica comum $\varphi(t)$ de $Y_1, \ldots, Y_n$ satisfaz

$$\int_{-\infty}^{+\infty} |\varphi(t)|^k dt < \infty \,. \tag{2.14}$$

O teorema seguinte é um corolário do teorema de Lindeberg-Lévy.

Teorema de DeMoivre-Laplace

Se $S_n \sim B(n,p)$ então $S_n^* = \frac{S_n - np}{\sqrt{np(1-p)}}$ tem distribuição normal $N(0,1)$ assintótica. Além disso, se $k = k_n$ depende de n mas $|(k-np)/\sqrt{np(1-p)}|$ permanece limitado quando $n \to \infty$, então

$$P(S_n = k) \sim \frac{1}{\sqrt{np(1-p)}} \phi\left(\frac{k-np}{\sqrt{np(1-p)}}\right), \tag{2.15}$$

com a notação $a_n \sim b_n$ significando que $\lim_{n\to\infty} \frac{a_n}{b_n} = 1$.

A equação (2.15) pode ser demonstrada por simples expansão de Taylor e aproximando os fatoriais do coeficiente binomial pela fórmula de Stirling (Seção 3.5, exemplo 3.7). A proporção de sucessos em n ensaios $\overline{Y}_n = S_n/n$ tem, portanto, uma distribuição normal $N(p, p(1-p)/n)$ assintótica implicando a fórmula aproximada

$$P(y_1 < \overline{Y}_n < y_2) \doteq \Phi(z_2) - \Phi(z_1) \,,$$

onde $z_i = (y_i - p)\sqrt{\frac{n}{p(1-p)}}$ para $i = 1, 2$.

O teorema de Lindeberg-Lévy é um caso especial do teorema seguinte mais geral.

Teorema Central do Limite

Seja $\{Y_n\}$ uma seqüência de variáveis aleatórias independentes (mas não necessariamente identicamente distribuídas) com os dois primeiros momentos $E(Y_i) = \mu_i$ e $\mathrm{Var}(Y_i) = \sigma_i^2$

finitos para $i = 1, 2, \ldots$ e com pelo menos um $\sigma_i^2 > 0$. Segundo condições gerais, tem-se

$$S_n^* = \frac{\sum_{i=1}^{n}(Y_i - \mu_i)}{\left(\sum_{i=1}^{n}\sigma_i^2\right)^{1/2}} \xrightarrow{\mathcal{D}} N(0,1) \; . \tag{2.16}$$

Várias condições que garantem a convergência em distribuição de S_n^* para a distribuição normal reduzida no teorema acima têm sido dadas por diferentes autores, incluindo generalizações para o caso de somas de variáveis aleatórias dependentes. No caso de variáveis independentes apresenta-se a seguir uma condição suficiente (teorema de Liapunov) e uma condição necessária e suficiente (teorema de Lindeberg-Feller) para que a convergência (2.16) seja satisfeita. Outras condições que garantem (2.16) estão fora do objetivo deste trabalho.

Teorema de Liapunov

Se para variáveis aleatórias independentes a relação

$$\lim_{n\to\infty} \frac{\left\{\sum_{i=1}^{n} E(|Y_i - \mu_i|^3)\right\}^{1/3}}{\left\{\sum_{i=1}^{n}\sigma_i^2\right\}^{1/2}} = 0$$

é satisfeita, então segue-se (2.16).

Teorema de Lindeberg-Feller

Suponha que para variáveis aleatórias independentes, $F_i(y)$ é a função de distribuição de Y_i e que $s_n^2 = \mathrm{Var}(S_n) = \sum_{i=1}^{n}\sigma_i^2$ satisfaz $\frac{\sigma_n^2}{s_n^2} \longrightarrow 0, s_n \longrightarrow \infty$ quando $n \longrightarrow \infty$. A convergência (2.16) é satisfeita se, e somente se, para todo $\epsilon > 0$

$$\lim_{n\to\infty} \frac{1}{s_n^2}\sum_{i=1}^{n}\int_{|y-\mu_i|>\epsilon s_n} (y-\mu_i)^2 dF_i(y) = 0 \; .$$

Uma conseqüência importante do teorema acima estabelece a seguinte condição: se para algum $k > 2$

$$\sum_{i=1}^{n} E(|Y_i - \mu_i|^k) = o(s_n^k)$$

quando $n \longrightarrow \infty$, então (2.16) é satisfeita.

Finalmente, torna-se de interesse prático e teórico caracterizar o erro do teorema central do limite, i.e., o erro da aproximação de $F_{S_n^*}(y) = P(S_n^* \leq y)$ por $\Phi(y)$. No caso *iid* tem-se a desigualdade de Berry-Esséen

$$\sup_y |F_{S_n^*}(y) - \Phi(y)| \leq \frac{33}{4} \frac{E(|Y_i - \mu|^3)}{\sqrt{n}\sigma^3}$$

que é válida para todo n e implica que a taxa de convergência de (2.12) é $n^{-1/2}$. Sob condições mais restritivas na expansão assintótica de $F_{S_n^*}(y) - \Phi(y)$ em potências de $1/\sqrt{n}$, pode ser demonstrado que (Ibragimov e Linnik, 1971)

$$|F_{S_n^*}(y) - \Phi(y)| \leq \frac{E\{|Y_i - \mu|^3\}}{\sigma^3\sqrt{2\pi\, n}}(1 - y^2)e^{-y^2/2} + o(n^{-1/2})$$

uniformemente em y.

2.6 Transformação Funcional

Um resultado muito útil de *transformação funcional* se refere ao comportamento assintótico de uma função de duas variáveis aleatórias, onde uma delas admite convergência em distribuição, não se impondo qualquer restrição sobre uma possível dependência entre essas variáveis aleatórias. Seja $h(Y_n, U_n)$ uma transformação funcional envolvendo duas variáveis aleatórias Y_n e U_n supondo que $Y_n \xrightarrow{\mathcal{D}} Y$ e $U_n \xrightarrow{P} k$, onde Y tem distribuição não-degenerada e k é uma constante finita. Admitindo-se que $h(y, u)$ é uma função contínua de u em $u = k$ para todos os pontos y no suporte de Y, pode-se demonstrar que $h(Y_n, U_n) \xrightarrow{\mathcal{D}} h(Y, k)$. Este resultado tem grande aplicabilidade na determinação de inúmeras distribuições assintóticas de funções de variáveis aleatórias. Em especial, $Y_n + U_n \xrightarrow{\mathcal{D}} Y + k$, $Y_n U_n \xrightarrow{\mathcal{D}} kY$ e $Y_n/U_n \xrightarrow{\mathcal{D}} Y/k$ se $k \neq 0$. Como motivação prática, suponha a estatística $T_n = \sqrt{n}(\overline{Y}_n - \mu)/s$ definida a partir de n variáveis aleatórias

$Y_1, \ldots, Y_n$ *iid* com média μ e variância σ^2, onde $\overline{Y}_n = \sum_{i=1}^{n} Y_i/n$ e $s_n^2 = \sum_{i=1}^{n}(Y_i - \overline{Y}_n)^2/(n-1)$. A distribuição exata de T_n é t_{n-1} (t de Student com $n-1$ graus de liberdade). Tem-se $E(s_n^2) = \sigma^2$ e $\lim_{n\to\infty} \mathrm{Var}(s_n^2) = 0$, de modo que $s_n^2 \xrightarrow{P} \sigma^2$ e, portanto, $s_n \xrightarrow{P} \sigma$. Pelo teorema central do limite $\sqrt{n}(\overline{Y}_n - \mu) \xrightarrow{\mathcal{D}} N(0, \sigma^2)$. Combinando as duas convergências obtém-se $T_n \xrightarrow{\mathcal{D}} N(0,1)$, resultado bastante conhecido de convergência da distribuição t de Student para a distribuição normal reduzida quando seus graus de liberdade tendem a infinito.

Uma situação comum na prática envolve a seqüência $\{Y_n\}$ admitindo-se as convergências $Y_n \xrightarrow{P} \mu$ e $\sqrt{n}(Y_n - \mu) \xrightarrow{\mathcal{D}} Y$, onde Y tem função de distribuição F arbitrária. Logo, $\sqrt{n}(Y_n - \mu) = Y + o_p(1)$ e

$$Y_n = \mu + \frac{Y}{\sqrt{n}} + o_p(n^{-1/2}) \, .$$

Em muitos casos, F é a função de distribuição Φ da normal reduzida.

Seja $\{h(Y_n)\}$ uma transformação funcional de $\{Y_n\}$, sendo $h(\cdot)$ uma função qualquer duas vezes diferenciável com $h'(\mu) \neq 0$ e $h''(y)$ suposta uniformente limitada no suporte de $\{Y_n\}$ para $n > n_0$. Por expansão de Taylor vem

$$\begin{aligned} \sqrt{n}\{h(Y_n) - h(\mu)\} &= \sqrt{n}h'(\mu)(Y_n - \mu) \\ &+ \frac{1}{2}\sqrt{n}h''(Z_n)(Y_n - \mu)^2 \, , \end{aligned} \tag{2.17}$$

onde $Z_n = \xi\mu + (1-\xi)Y_n$ para $\xi \in (0,1)$. Como h'' é limitada, o segundo termo em (2.17) é $O_p(n^{-1/2})$. Assim, a equação de linearização decorre de (2.17)

$$\sqrt{n}\{h(Y_n) - h(\mu)\} = \sqrt{n}h'(\mu)(Y_n - \mu) + o_p(1) \, . \tag{2.18}$$

Por hipótese $\sqrt{n}h'(\mu)(Y_n - \mu) \xrightarrow{\mathcal{D}} h'(\mu)Y$ e, então, (2.18) implica que

$$\frac{\sqrt{n}\{h(Y_n) - h(\mu)\}}{h'(\mu)} \xrightarrow{\mathcal{D}} Y \, . \tag{2.19}$$

Estimando-se $h'(\mu)$ por $h'(Y_n)$ segue, também, a convergência

$$\frac{\sqrt{n}\{h(Y_n) - h(\mu)\}}{h'(Y_n)} \xrightarrow{\mathcal{D}} Y .$$

Em especial, se $Y \sim N(0, \sigma^2)$, então (2.19) conduz ao resultado

$$\sqrt{n}\{h(Y_n) - h(\mu)\} \xrightarrow{\mathcal{D}} N(0, h'(\mu)^2\sigma^2) .$$

Além disso, se $\sigma = \sigma(\mu)$ é uma função contínua de μ, sendo estimada por $\sigma(Y_n)$, obtém-se também,

$$\frac{\sqrt{n}\{h(Y_n) - h(\mu)\}}{\sigma(Y_n)h'(Y_n)} \xrightarrow{\mathcal{D}} N(0,1) .$$

Exemplo 2.7 *Supõe-se que* $\sqrt{n}(Y_n - \mu) \xrightarrow{\mathcal{D}} N(0, \sigma^2)$ *e sejam* $h_1(Y_n) = Y_n^2$ *e* $h_2(Y_n) = \sqrt{Y_n}$. *Então,* $\sqrt{n}(Y_n^2 - \mu^2) \xrightarrow{\mathcal{D}} N(0, 4\mu^2\sigma^2)$ *e* $\sqrt{n}(\sqrt{Y_n} - \sqrt{\mu}) \xrightarrow{\mathcal{D}} N(0, \sigma^2/(4\mu))$.

Os momentos centrais definidos a partir de n observações *iid* $Y_1, \ldots, Y_n$ por $m_k = n^{-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^k$ $(k = 1, 2, \ldots)$ são freqüentes nas fórmulas das estatísticas e é importante conhecer suas propriedades em grandes amostras. Para $k = 1$ e $k = 2$ tem-se a média e a variância amostrais. Pode-se demonstrar que (vide Serfling, 1980, Seção 2.2.3)

(i) $m_k \xrightarrow{q.c.} \mu_k$;

(ii) o viés de m_k é dado por

$$E(m_k - \mu_k) = \frac{k(k-1)\mu_{k-1}\mu_2 - 2k\mu_k}{2n} + O(n^{-2});$$

(iii) a variância de m_k iguala $\mathrm{Var}(m_k) = \frac{\alpha_k}{n} + O(n^{-2})$, onde

$$\alpha_k = \mu_{2k} - \mu_k^2 - 2k\mu_{k-1}\mu_{k+1} + k^2\mu_2\mu_{k-1}^2;$$

(iv) $\sqrt{n}(m_k - \mu_k) \xrightarrow{\mathcal{D}} N(0, \alpha_k)$ com α_k dado em (iii).

Os resultados (i) – (iv) são verdadeiros para qualquer distribuição de Y. Note-se que a média e a variância de m_k estão definidas em termos dos momentos centrais $\mu_2, \mu_{k-1}, \mu_k, \mu_{k+1}$ e μ_{2k} de Y. O item (iv) para $k = 1$ e 2 produz

$$\sqrt{n}\,\overline{Y} \xrightarrow{\mathcal{D}} N(0, \sigma^2) \text{ e } \sqrt{n}s^2 \xrightarrow{\mathcal{D}} N(0, \mu_4 - \sigma^4)\ ,$$

pois $\mu_1 = 0$ e $\mu_2 = \sigma^2$. Portanto,

$$\sqrt{n}s \xrightarrow{\mathcal{D}} N(0, (\mu_4 - \sigma^4)/(4\sigma^4))\ .$$

A equação (2.18) escrita como

$$h(Y_n) = h(\mu) + h'(\mu)(Y_n - \mu) + o_p(n^{-1/2})$$

pode ser generalizada, supondo que $h(\cdot)$ é uma função real diferenciável até ordem k, para

$$h(Y_n) = \sum_{j=0}^{k} \frac{h^{(j)}(\mu)}{j!}(Y_n - \mu)^j + o_p(n^{-j/2})\ .$$

Os momentos (e, então, os cumulantes) de $h(Y_n)$ até uma ordem pré-fixada podem ser obtidos a partir dos momentos de Y_n elevando-se a expansão acima a potências de ordens dos momentos a serem calculados.

Finaliza-se este capítulo tratando o problema de *estabilização da variância* na estimação de um parâmetro θ através de uma estatística Y_n que é assintoticamente normal mas sua variância assintótica depende de θ. Suponha que $\sqrt{n}(Y_n - \theta)/v(\theta)^{1/2} \xrightarrow{\mathcal{D}} N(0,1)$, ou seja, $v(\theta)/n$ é a variância assintótica de Y_n. Neste caso, a região de rejeição do parâmetro θ depende de θ através de $v(\theta)$ e pode não haver a propriedade desejável de monotonicidade no parâmetro. Objetiva-se determinar uma transformação $h(Y_n)$ para se fazer inferência sobre $\tau = h(\theta)$ de modo que

$$\frac{\sqrt{n}(h(Y_n) - h(\theta))}{k} \xrightarrow{\mathcal{D}} N(0,1)\ ,$$

ou seja, a variância assintótica k^2/n de $h(Y_n)$ é uma constante independente de θ. Tem-se

$$\mathrm{Var}(h(Y_n)) = h'(\theta)^2 \mathrm{Var}(Y_n) .$$

Então, $k^2 = h'(\theta)^2 v(\theta)$, implicando

$$h(\theta) = k \int_0^t \frac{dt}{\sqrt{v(t)}} . \tag{2.20}$$

Dado $v(\theta)$ obtém-se de (2.20) a transformação estabilizadora e o intervalo de confiança segue baseado em $h(Y_n)$ e $\tau = h(\theta)$, i.e., $\sqrt{n}|h(Y_n) - \tau| \leq k z_{\alpha/2}$, onde $z_{\alpha/2}$ é o ponto crítico da distribuição $N(0,1)$ correspondente ao nível de significância α. Por exemplo, se $v(\theta) = \theta^{2m}$ e $m \neq 1$ vem $h(\theta) = k\theta^{1-m}/(1-m)$. Para $m = -1, 2$ e $1/2, h(\theta)$ iguala $k\theta^2/2, -k/\theta$ e $2k\sqrt{\theta}$, respectivamente. Para $m = 1, h(\theta) = k \log \theta$. Cada um desses valores de m corresponde a uma distribuição importante. Sejam os casos $m = 1/2$ e $m = 1$. O primeiro caso pode ser caracterizado pela soma S_n de n variáveis aleatórias *iid* com distribuição de Poisson $P(\theta)$. Logo, $S_n^* = \sqrt{n}(\overline{Y}_n - \theta)/\sqrt{\theta} \xrightarrow{\mathcal{D}} N(0,1)$ e $v(\theta) = \theta$. Assim, $h(\theta) = 2k\sqrt{\theta}$ e a variância da raiz quadrada de $\overline{Y}_n$ é estabilizada, ou seja, $\sqrt{n}(\sqrt{\overline{Y}_n} - \sqrt{\theta}) \xrightarrow{\mathcal{D}} N(0,1)$. O segundo caso ($m = 1$) pode ser exemplificado pela soma S_n de n variáveis aleatórias *iid* com distribuição gama parametrizada pela média θ e pelo parâmetro de forma p (vide, também, exemplo 2.1). Assim, $E(Y) = \theta$ e $\mathrm{Var}(\theta) = \theta^2/p$. A soma S_n^* padronizada é $S_n^* = \sqrt{np}(\overline{Y}_n - \theta)/\theta$ e $v(\theta) = \theta^2/p$. Então, $h(\theta) = k\sqrt{p} \log \theta$ e a variância de $h(\overline{Y}_n)$ é estabilizada mediante a transformação logarítmica. Tem-se,

$$\sqrt{np}(\log \overline{Y}_n - \log \theta) \xrightarrow{\mathcal{D}} N(0,1) .$$

2.7 Exercícios

1. Mostre que a variável qui-quadrado padronizada $(\chi_n^2 - n)/\sqrt{2n}$ converge em distribuição para a normal $N(0,1)$. Avalie um limite para o erro desta aproximação.

2. Mostre que a variável aleatória Y com função densidade $f(y) = \{\cosh(\pi y)\}^{-1}$, $y \in \mathbb{R}$, tem função característica $\varphi(t) = \mathrm{sech}(t/2)$.

3. Demonstre que os momentos ordinários da variável aleatória com função densidade $f(y) = ky^p e^{-\gamma/y}$, $\gamma > 0$, $y \geq 0$ são $\mu'_r = \gamma^r \Gamma(p-1-r)/\Gamma(p-1)$ se $r < p-1$ e que, caso contrário, os momentos não existem.

4. Justifique que a distribuição $N(\mu, \sigma^2)$ é determinada univocamente pelos seus momentos.

5. Mostre que: (a) a distribuição exponencial cuja função densidade é dada por $f(y) = \sigma^{-1} e^{-y/\sigma} (\sigma > 0)$ tem cumulantes $\kappa_r = \sigma^r (r-1)!$, $r = 1, 2, \ldots$; (b) a função $\exp(-t^\alpha)$ não pode ser uma função característica, exceto se $\alpha = 2$.

6. Mostre que: (a) se $Y \xrightarrow{\mathcal{D}} Y$, então $Y_n = O_p(1)$; (b) se $Y_n = o_p(U_n)$, então $Y_n = o_p(U_n)$; (c) se $Y_n \xrightarrow{\mathcal{D}} Y$ e $X_n \xrightarrow{P} X$, então $Y_n + X_n \xrightarrow{\mathcal{D}} Y + X$.

7. Seja $\sigma_n^{-1}(Y_n - \mu) \xrightarrow{\mathcal{D}} N(0,1)$. Então, $Y_n \xrightarrow{P} \mu$ se, e somente se, $\sigma_n \longrightarrow 0$ quando $n \longrightarrow \infty$.

8. Seja $\varphi(t)$ a função característica da variável aleatória Y. Mostre que se Y é contínua $\lim_{|t|\to\infty} \varphi(t) = 0$ e se Y é discreta $\lim_{|t|\to\infty} \sup |\varphi(t)| = 1$.

9. Suponha as convergências $\sqrt{n}(Y_n - \mu)/\sigma \xrightarrow{\mathcal{D}} N(0,1)$ e $\sqrt{n}(X_n - c)/\sqrt{v} \xrightarrow{\mathcal{D}} N(0,1)$, $c \neq 0$. Mostre que $c\sqrt{n}(Y_n - \mu)/(\sigma X_n) \xrightarrow{\mathcal{D}} N(0,1)$.

10. Demonstre que as funções características das distribuições logística e de Laplace cujas funções densidades são $f(y) = e^{-y}(1+e^{-y})^{-2}$ e $f(y) = \exp\{-|y-\mu|/\sigma\}/(2\sigma)$, $y \in \mathbb{R}$ em ambos os casos, são dadas por $\varphi(t) = \Gamma(1-it)\Gamma(1+it)$ e $\varphi(t) = e^{it\mu}(1+\sigma^2 t^2)^{-1}$, respectivamente.

11. Sejam $(Y_n - \mu)/\sigma_n \xrightarrow{\mathcal{D}} N(0,1)$ e $X_n = 0$ e n com probabilidades $1 - n^{-1}$ e n^{-1} respectivamente. Mostre que $(Y_n + X_n - \mu)/\sigma_n \xrightarrow{\mathcal{D}} N(0,1)$.

12. Mostre que se $\varphi(t)$ é a função característica de uma variável aleatória, $\varphi(t)^2$ também é uma função característica.

13. (a) Uma variável aleatória tem momentos $\mu'_r = k/(k+r)$, $r = 1, 2, \ldots$, onde $k > 0$. Mostre que sua função densidade é $f(y) = y^{k-1}$, $y \in (0,1)$; (b) uma variável

aleatória tem função característica $\varphi(t) = (1+t^2)^{-1}$. Mostre que sua função densidade é $f(y) = e^{-|y|}/2,\ y \in \mathbb{R}$.

14. Se Y é uma variável aleatória tendo momentos $\mu'_r = (k+r)!/k!,\ k$ um inteiro positivo, então a sua função densidade é univocamente determinada por $f(y) = y^k e^{-y}/k!,\ y > 0$.

15. Se $Y_1, \ldots, Y_n$ satisfazem às suposições do teorema de Lindeberg-Lévy e, além disso, o momento $E(|Y_i|^3)$ existe, então $S_n^* = \sqrt{n}(\overline{Y}_n - \mu)/n$ tem *fda* que satisfaz

$$|F_{S_n^*}(y) - \Phi(y)| \leq \frac{k\,E(|Y_i|^3)}{\sqrt{n}\sigma^3},$$

onde k é uma constante.

16. Se Y é uma variável aleatória tal que $E(e^{kY})$ existe para $k > 0$, então

$$P(Y \geq \epsilon) \leq E(e^{kY})/e^{k\epsilon}\ .$$

17. Se $Y_1, Y_2, \ldots$ é uma seqüência de variáveis aleatórias *iid.* Se $E(Y_i) = \mu$ é finito, então $\overline{Y}_n \xrightarrow{P} \mu$.

18. A função densidade da distribuição de Laplace tem a forma $f(y;\mu,\phi) = (2\phi)^{-1}\exp(-|y-\mu|/\phi),\ \phi > 0$. Mostre que a sua função característica é dada por $\varphi(t) = (1+\phi^2 t^2)^{-1}\exp(it\mu)$. Mostre que ela tem momentos de todas as ordens e que não é preservada segundo convolução.

Capítulo 3

Expansões Assintóticas

3.1 Introdução

Considere uma *expansão assintótica* para a função $g_n(y)$ em algum ponto fixo y expressa para $n \to \infty$ como

$$g_n(y) = f(y)\left\{1 + \frac{\gamma_1(y)}{\sqrt{n}} + \frac{\gamma_2(y)}{n} + \frac{\gamma_3(y)}{n\sqrt{n}} + \cdots\right\}, \tag{3.1}$$

onde n é usualmente o tamanho da amostra ou uma quantidade de informação. Na inferência a função $g_n(y)$ de interesse é tipicamente uma função densidade (ou de distribuição) de uma estatística baseada numa amostra de tamanho n e $f(y)$ pode ser considerada uma aproximação de primeira ordem tal qual a função densidade (ou de distribuição) da normal reduzida. A função $g_n(y)$ pode ser definida diretamente de uma seqüência de comprimento n de variáveis aleatórias, por exemplo, como a função densidade da média amostral $\overline{Y}_n = n^{-1}\sum_{i=1}^{n} Y_i$ de n variáveis aleatórias *iid* sendo $f(y)$ sua função densidade limite, que é usualmente a função densidade $\phi(y)$ da normal reduzida. Ela pode ser também uma função geratriz de momentos ou cumulantes. Embora a equação (3.1) seja definida para um valor fixo y, tem-se o interesse em saber para qual região dos valores de y ela permanece válida como uma expansão assintótica.

Uma característica importante da expansão assintótica (3.1) é que ela não é, em geral, uma série convergente para $g_n(y)$ e, assim, tomando-se mais termos no seu lado direito a aproximação para $g_n(y)$ não necessariamente melhora.

As expansões assintóticas são usadas rotineiramente em muitas áreas da análise matemática. Os livros de Jeffreys (1962), DeBruijn (1970) e Bleistein e Handelsman (1975) são excelentes fontes para estudos aprofundados. Embora uma aproximação do tipo (3.1) seja somente válida quando $n \longrightarrow \infty$, obtém-se frequentemente uma boa precisão mesmo para valores pequenos de n. Há interesse de investigar em cada caso a precisão da aproximação (3.1) para vários valores de y, bem como o intervalo de variação de y para o qual o erro da aproximação é uniforme. Apresentam-se a seguir expansões do tipo (3.1) permitindo o termo principal $f(y)$ depender de n para algumas funções matemáticas de grande interesse na Estatística:

(i) A função gama $\Gamma(n) = \int_0^\infty x^{n-1}e^{-x}dx$ admite para n grande a expansão de Stirling expressa somente em potências de n^{-1}

$$\Gamma(n) = (2\pi)^{1/2}e^{-n}n^{n-0,5}\left\{1 + \frac{1}{12n} + \frac{1}{288n^2} - \frac{139}{51840n^3} - \frac{571}{2488320n^4} + O(n^{-5})\right\}.$$

Fixando n e tomando mais termos no lado direito da fórmula acima, o erro da aproximação aumenta. Para valores de $n \geq 5$, a fórmula assintótica $\sqrt{2\pi}\, e^{-n}n^{n-0,5}$ é suficiente para muitos propósitos;

(ii) A função gama incompleta $\Gamma(k, y) = \int_y^\infty e^{-t}t^{k-1}dt$ admite a expansão

$$\Gamma(k, n) = n^{k-1}e^{-n}\left\{1 + \frac{k-1}{n} + \frac{(k-1)(k-2)}{n^2} + O(n^{-3})\right\};$$

(iii) A função $\log y - \psi(y)$, onde $\psi(y) = \frac{d\log\Gamma(y)}{dy}$ é a função digama, é estudada na estimação do parâmetro de forma da distribuição gama. Valores inteiros são computados como $\psi(1) = -\gamma$, $\psi(n) = -\gamma + \sum_{k=1}^{n-1} k^{-1} (n \geq 2)$, onde $\gamma = 0,5772156649\ldots$ é a constante de Euler. Tem-se a expansão, quando o argumento $y \longrightarrow \infty$ (ao invés de $n \longrightarrow \infty$)

$$\log y - \psi(y) = \frac{1}{2y}\left\{1 + \frac{1}{6y} - \frac{1}{60y^3} + \frac{1}{126y^5} + O(y^{-7})\right\}.$$

Neste capítulo são apresentadas várias expansões importantes do tipo (3.1), geral-

mente até termos de ordem n^{-1}. Entre estas expansões, citam-se as expansões de Gram-Charlier, Edgeworth, Cornish-Fisher, ponto de sela, Laplace e as expansões que relacionam funções de distribuição e de variáveis aleatórias. O leitor que desejar maiores detalhes matemáticos poderá consultar os livros de McCullagh (1987, Capítulos 5 e 6), Barndorff-Nielsen e Cox (1990, Capítulo 4) e Hinkley, Reid e Snell (1991, Capítulo 12).

3.2 Expansão de Gram-Charlier

Seja $f(y)$ uma função densidade conhecida, cujos cumulantes são dados por $\kappa_1, \kappa_2, \ldots$ O interesse reside em usar $f(y)$ para aproximar uma função densidade $g(y)$ (em geral desconhecida) a partir da aplicação de um operador $T(D)$ a $f(y)$. O operador é formulado como $T(D) = \exp\left\{\sum_{j=1}^{\infty} \epsilon_j (-D)^j / j!\right\}$ e a aproximação para $g(y)$ é definida por

$$g(y) = T(D) f(y) \, ,$$

onde D é o operador diferencial, ou seja, $D^j f(y) = d^j f(y)/dy^j$.

Os cumulantes de $g(y)$ são determinados como os coeficientes de $t^r/r!$ na expansão de $\log\left\{\int_{-\infty}^{+\infty} e^{ty} g(y) dy\right\}$ (Seção 2.3). Expandindo o operador $T(D)$ em série de Taylor vem $T(D) = \sum_{i=0}^{\infty} \frac{1}{i!} \sum_{j=1}^{\infty} \frac{\epsilon_j (-D)^j}{j!}$ de onde se conclui que os cumulantes de $g(y)$ são dadas por $\kappa_1 + \epsilon_1, \kappa_2 + \epsilon_2, \ldots$ A função $g(y)$ pode não satisfazer a condição $g(y) \geq 0$ para todo y, mas seus cumulantes $\kappa_r + \epsilon_r$ são definidos mesmo que esta condição não seja satisfeita.

De $g(y) = T(D) f(y)$ obtém-se, pela expansão de $T(D)$,

$$\begin{aligned} g(y) \;=\; & f(y) - \epsilon_1 D f(y) + \frac{1}{2}(\epsilon_1^2 + \epsilon_2) D^2 f(y) \\ & - \frac{1}{6}(\epsilon_1^3 + 3\epsilon_1\epsilon_2 + \epsilon_3) D^3 f(y) + \frac{1}{24}(\epsilon_1^4 + 6\epsilon_1^2\epsilon_2 + 4\epsilon_1\epsilon_3 + \epsilon_4) D^4 f(y) + \cdots \end{aligned} \tag{3.2}$$

A equação (3.2) mostra que a função densidade $g(y)$ de uma variável aleatória contínua qualquer pode ser expandida em termos de uma função densidade $f(y)$ de referência conhecida e de suas derivadas, cujos coeficientes são funções de diferenças ($\epsilon_i's$) entre os cumulantes correspondentes associados às funções densidade $g(y)$ e $f(y)$. Em muitos

casos, $D^j f(y) = P_j(y)f(y)$, onde $P_j(y)$ é um polinômio de grau j em y. Esses polinômios são geralmente *ortogonais* com relação à distribuição associada a $f(y)$, ou seja, $\int P_j(y)P_k(y)f(y) = 0$ para $j \neq k$. Por exemplo, se $f(y)$ é a função densidade $\phi(y)$ da distribuição normal reduzida, $(-1)^j P_j(y)$ é o polinômio de Hermite $H_j(y)$ de grau j definido pela identidade

$$(-D)^r \phi(y) = H_r(y)\phi(y) .$$

Os primeiros polinômios de Hermite são $H_0(y) = 1$, $H_1(y) = y$, $H_2(y) = y^2 - 1$, $H_3(y) = y^3 - 3y$, $H_4(y) = y^4 - 6y^2 + 3$, $H_5(y) = y^5 - 10y^3 + 15y$ e $H_6(y) = y^6 - 15y^4 + 45y^2 - 15$. Esses polinômios têm propriedades interessantes decorrentes da identidade

$$\exp(ty - t^2/2) = \sum_{j=0}^{\infty} \frac{t^j}{j!} H_j(y) ,$$

tais como:

$$\frac{d}{dy} H_r(y) = r\, H_{r-1}(y) ,$$

$$D^j H_r(y) = r^{(j)} H_{r-j}(y) \quad \text{para} \quad r \geq j,$$

onde $r^{(j)} = r(r-1)\cdots(r-j+1)$. Satisfazem ainda a relação de recorrência

$$H_r(y) = yH_{r-1}(y) - (r-1)H_{r-2}(y) \quad (r \geq 2) .$$

Suponha agora que as médias e as variâncias de $g(y)$ e $f(y)$ são tomadas iguais, por exemplo, pela padronização através de transformação linear das variáveis. Neste caso, $\epsilon_1 = \epsilon_2 = 0$ e (3.2) implica

$$g(y) = f(y) - \left\{ \frac{\epsilon_3}{3!} P_3(y) - \frac{\epsilon_4}{4!} P_4(y) + \cdots \right\} f(y) . \tag{3.3}$$

Integrando (3.3) obtém-se uma relação equivalente para as funções de distribuição $G(y) = \int_{-\infty}^{y} g(t)dt$ e $F(y) = \int_{-\infty}^{y} f(t)dt$ correspondentes a $g(y)$ e $f(y)$:

$$G(y) = F(y) - \left\{ \frac{\epsilon_3}{3!} P_2(y) - \frac{\epsilon_4}{4!} P_3(y) + \cdots \right\} f(y) . \tag{3.4}$$

O caso especial mais importante e de maior aplicabilidade das expansões (3.3) e (3.4) surge quando $f(y)$ é a função densidade $\phi(y)$ da distribuição normal reduzida. Neste

caso, $\kappa_r = 0$ para $r > 2$ (Seção 2.3) e $\epsilon_3, \epsilon_4, \ldots$ se igualam aos cumulantes de $g(y)$. Assim, (3.3) simplifica-se para

$$g(y) = \phi(y)\left\{1 + \frac{\epsilon_3}{3!}H_3(y) + \frac{\epsilon_4}{4!}H_4(y) + \frac{\epsilon_5}{5!}H_5(y) + \frac{(\epsilon_6 + 10\epsilon_3^2)}{6!}H_6(y) + \cdots\right\}. \quad (3.5)$$

A expansão (3.5) é denominada *expansão de Gram-Charlier*. Usualmente, não se consideram em (3.5) polinômios de ordem superior a seis. Os termos em (3.5) ocorrem numa seqüência determinada pelas derivadas sucessivas de $\phi(y)$. Entretanto, esta seqüência não se apresenta necessariamente em ordem decrescente de magnitude e, algumas vezes, uma ordenação diferente deve ser adotada a partir da avaliação da magnitude dos seus vários termos. Integrando (3.5) e usando a relação $\int H_r(y)\phi(y)dy = -H_{r-1}(y)\phi(y)(r \geq 1)$ vem

$$G(y) = \Phi(y) - \left\{\frac{\epsilon_3}{3!}H_2(y) + \frac{\epsilon_4}{4!}H_3(y) + \frac{\epsilon_5}{5!}H_4(y) + \frac{(\epsilon_6 + 10\epsilon_3^2)}{6!}H_5(y) + \cdots\right\}\phi(y), \quad (3.6)$$

onde $\Phi(y)$ é a função de distribuição da normal reduzida.

As fórmulas (3.5) e (3.6) mostram que as funções densidade e de distribuição de uma variável aleatória qualquer Y padronizada podem, em geral, ser expressas por expansões envolvendo seus cumulantes, os polinômios de Hermite e as funções densidade e de distribuição da normal reduzida. Nas aplicações de (3.5) e (3.6) é importante coletar os termos de mesma magnitude, conforme mostra o exemplo seguinte.

Exemplo 3.1 *Seja Z uma variável aleatória com distribuição gama de parâmetros 1(escala) e $\lambda > 0$ (forma). A função geratriz de cumulantes de Z é $K(t) = -\lambda\log(1-t)$ e os seus cumulantes igualam $\gamma_r = \lambda(r-1)!$. Deseja-se obter uma aproximação para a função densidade $g(y)$ da variável gama padronizada $Y = (Z - \gamma_1)/\gamma_2^{1/2}$ em termos da função densidade $\phi(y)$ da distribuição normal reduzida. Os cumulantes de Y são dados por $\kappa_r + \epsilon_r = (r-1)!\lambda^{(2-r)/2}$ (vide Seção 2.3), sendo $\kappa_1 = 0$, $\kappa_2 = 1$ e $\kappa_r = 0$, $r > 2$. Na expansão de Gram-Charlier da função densidade $g(y) = \phi(y)\sum_{j=0}^{12} c_j H_j(y)$, decorrente de (3.5) e até o termo envolvendo $H_{12}(y)$, os coeficientes c_j têm as seguintes ordens de magnitude em λ:*

c_0	c_3	c_4	c_5	c_6	c_7	c_8	c_9	c_{10}	c_{11}	c_{12}
0	$-1/2$	-1	$-3/2$	-1	$-3/2$	-2	$-3/2$	-2	$-5/2$	-2

Assim, os termos da expansão de Gram-Charlier não necessariamente decrescem em ordem de magnitude de λ. Deve-se ter cuidado ao truncar (3.5) de modo que todos os termos não incluídos sejam realmente de ordem inferior àqueles da expansão truncada. Por exemplo, para obter uma expansão corrigida para $g(y)$ até ordem $\lambda^{-3/2}$, somente os termos correspondentes a c_8, c_{10}, c_{11} e c_{12} não seriam incluídos.

3.3 Expansões de Edgeworth

Trata-se aqui das expansões de Edgeworth para somas padronizadas de variáveis aleatórias univariadas *iid*. Estas expansões são importantes na teoria assintótica quando a integral de convolução referente à soma de variáveis aleatórias não pode ser calculada explicitamente. A extensão para o caso de variáveis multivariadas está fora dos objetivos deste texto e o leitor poderá consultar o livro de McCullagh (1987, Capítulo 5).

Seja Y uma variável aleatória com funções densidade $f(y)$ e geratriz de cumulantes $K(t)$. Os cumulantes padronizados de Y são $\rho_r = \kappa_r/\kappa_2^{r/2}$ para $r \geq 2$. Tem-se $\kappa_1 = E(y) = \mu$ e $\kappa_2 = \mathrm{Var}(Y) = \sigma^2$. Suponha que $Y_1, \ldots, Y_n$ são realizações *iid* de Y e sejam: $S_n = \sum_{i=1}^{n} Y_i$, a soma estocástica e $S_n^* = (S_n - n\mu)/(\sigma\sqrt{n})$, a soma padronizada. Como as variáveis aleatórias são *iid*, as funções geratrizes de cumulantes de S_n e S_n^* são dadas por $K_{S_n}(t) = nK(t)$ e

$$K_{S_n^*}(t) = -\frac{\sqrt{n}\mu t}{\sigma} + nK\left(\frac{t}{\sigma\sqrt{n}}\right), \tag{3.7}$$

respectivamente. A expansão de $K(t)$ em série de Taylor equivale a uma soma de funções dos cumulantes padronizados de Y

$$K(t) = \mu t + \sigma^2 t^2/2 + \rho_3\sigma^3 t^3/6 + \rho_4\sigma^4 t^4/24 + \cdots$$

que substituída em (3.7) implica

$$K_{S_n^*}(t) = t^2/2 + \rho_3 t^3/(6\sqrt{n}) + \rho_4 t^4/(24n) + O(n^{-3/2}). \tag{3.8}$$

A expansão (3.8) revela o esperado, ou seja, que $K_{S_n^*}(t) \longrightarrow t^2/2$ quando $n \longrightarrow \infty$, pois pelo teorema central do limite (Seção 2.5) S_n^* converge em distribuição para a distribuição normal $N(0,1)$ quando n tende a infinito. A função geratriz de momentos

$M_{S_n^*}(t)$ de S_n^* é obtida de (3.8) tomando exponenciais. Logo,

$$M_{S_n^*}(t) = \exp(t^2/2)\{1 + \rho_3 t^3/(6\sqrt{n}) + \rho_4 t^2/(24n) + \rho_3^2 t^6/(72n) + O(n^{-3/2})\}.$$

Para obter a função densidade de S_n^*, a equação acima deve ser invertida termo a termo usando a identidade

$$\int e^{ty}\phi(y)H_r(y)dy = t^r \exp(t^2/2) \ .$$

Então, a função densidade de S_n^* é dada por

$$f_{S_n^*}(y) = \phi(y)\left\{1 + \frac{\rho_3}{6\sqrt{n}}H_3(y) + \frac{\rho_4}{24n}H_4(y) + \frac{\rho_3^2}{72n}H_6(y)\right\} + O(n^{-3/2}) \ . \qquad (3.9)$$

A integral de (3.9) produz a expansão da função de distribuição de S_n^* como

$$F_{S_n^*}(y) = \Phi(y) - \phi(y)\left\{\frac{\rho_3}{6\sqrt{n}}H_2(y) + \frac{\rho_4}{24n}H_3(y) + \frac{\rho_3^2}{72n}H_5(y)\right\} + O(n^{-3/2}) \ . \qquad (3.10)$$

As fórmulas (3.9) e (3.10) são as *expansões de Edgeworth* para as funções densidade e de distribuição de uma soma padronizada S_n^*, respectivamente. É importante salientar que a expansão (3.9) segue diretamente da expansão de Gram-Charlier (3.5), pois os cumulantes de S_n^* são simplesmente $\epsilon_r = O(n^{1-r/2})$ para $r \geq 3$ com $\epsilon_3 = \rho_3/\sqrt{n}$ e $\epsilon_4 = \rho_4/n$. O termo principal em (3.10) é a função de distribuição $\Phi(y)$ da normal reduzida, como previsto pelo teorema central do limite. O termo de ordem $n^{-1/2}$ é um ajustamento face à assimetria da distribuição de Y e os termos de ordem n^{-1} representam um ajustamento simultâneo devido à assimetria e curtose da distribuição de Y.

A adequação das aproximações $\phi(y), \phi(y)\{1 + \rho_3 H_3(y)/(6\sqrt{n})\}$ e (3.9) para a função densidade de S_n^* depende do valor de y. A aproximação (3.9) poderá não ser apropriada nas extremidades da distribuição de S_n^* quando $|y|$ crescer, pois os polinômios de Hermite não são limitados. No ponto $y = 0$, o erro da aproximação normal $\phi(y)$ é $O(n^{-1})$ e não $O(n^{-1/2})$, enquanto o da expansão (3.9) é $O(n^{-2})$, pois os termos de potência ímpar em $n^{-1/2}$ dependem apenas de polinômios de grau ímpar e todos eles se anulam para $y = 0$. Assim, desejando-se aproximar a função densidade de S_n^* na origem, $f_{S_n^*}(0)$, obtém-se uma expansão em potências de n^{-1} ao invés de potências de $n^{-1/2}$. Quando $\rho_3 \neq 0$ (distribuições de Y assimétricas) o termo de ordem $n^{-1/2}$ poderá ser muito grande nas ex-

tremidades da distribuição de S_n^* quando $H_3(y)$ for apreciável, invalidando a aproximação $\phi(y)\{1+\rho_3 H_3(y)/(6\sqrt{n})\}$ para a função densidade de S_n^*. Diferentemente, a aproximação em torno da média $E(S_n^*)=0$, onde $H_3(0)=0$, será satisfatória, pois envolverá somente termos de ordem n^{-1}. Obviamente, se a função densidade de Y é simétrica ($\rho_3=0$), a aproximação normal usual para a função densidade de S_n^* estará correta até ordem $n^{-1/2}$ ao invés de até ordem 1.

A função $\phi(y)\{1+\rho_3 H_3(y)/(6\sqrt{n})\}$ formada pelos dois primeiros termos de (3.9) não é uma função densidade em y para n fixo e $\rho_3 \neq 0$, pois para $\rho_3 y$ suficientemente grande e negativo, o valor desta função pode ser negativo. Entretanto, isto não contradiz a suposição assintótica da validade de (3.9) que é "y fixado e $n \longrightarrow \infty$". Uma forma de superar esta dificuldade é escrever a aproximação acima como $\phi(y)\exp\{\rho_3 H_3(y)/(6\sqrt{n})\}$. Entretanto, esta forma tem a desvantagem de ser ilimitada e, portanto, pode não ser normalizada exatamente em $\mathbb{R}$.

O erro em (3.10) só será $O(n^{-3/2})$ se S_n^* tiver distribuição contínua. No caso discreto, a função de distribuição exata de S_n^* é descontínua nos seus possíveis valores, com saltos de ordem $O(n^{-1/2})$. A aproximação (3.10) é contínua e deve envolver erros de ordem $n^{-1/2}$ próximo aos pontos de descontinuidade. Entretanto, Kolassa e McCullagh (1990) propõem uma versão de (3.10), válida até $O(n^{-1})$ para distribuições discretas, pelo ajustamento dos cumulantes ρ_3 e ρ_4 através das correções de Sheppard.

Exemplo 3.2 *Sejam $Y_1,\ldots,Y_n$ variáveis aleatórias iid com distribuição exponencial de média um. A função densidade exata de S_n^* é dada por*

$$\pi_{S_n^*}(y)=\sqrt{n}(n+y\sqrt{n})^{n-1}\exp(-n-y\sqrt{n})/(n-1)!\ .$$

Para obter de (3.9) a expansão de Edgeworth tem-se $E(S_n)=n$, $\mathrm{Var}(S_n)=n$, $\rho_3=2$ e $\rho_4=6$. Logo,

$$f_{S_n^*}(y)=\phi(y)\left\{1+\frac{H_3(y)}{3\sqrt{n}}+\frac{H_4(y)}{4n}+\frac{H_6(y)}{18n}\right\}+O(n^{-3/2})\ .$$

Na Tabela 3.1 compara-se para $n=5$ o valor exato $\pi_{S_n^*}(y)$ com a aproximação normal $\phi(y)$ (termo principal) e com aquelas expansões de $f_{S_n^*}(y)$ obtidas da equação acima considerando apenas o termo $O(n^{-1/2})$ e com aqueles dois termos de ordem $O(n^{-1})$.

Tabela 3.1: *Aproximações de Edgeworth para a função densidade da soma padronizada de 5 variáveis exponenciais iid*

			Expansões de Edgeworth	
y	Exato	Normal	até $O(n^{-1/2})$	até $O(n^{-1})$
-2	0,0043	0,0540	0,0379	0,0178
-1,5	0,1319	0,1295	0,1512	0,1480
-1,0	0,3428	0,2420	0,3141	0,3329
-0,5	0,4361	0,3521	0,4242	0,4335
0	0,3924	0,3989	0,3989	0,3922
1	0,1840	0,2420	0,1698	0,1887
2	0,0577	0,0540	0,0701	0,0500
3	0,0144	0,0044	0,0163	0,0181

Para valores pequenos de n, a expansão de Edgeworth não é boa nas caudas da distribuição de S_n^*. A Tabela 3.1 mostra que, fora dessas caudas, a expansão de Edgeworth incluindo os termos $O(n^{-1})$ é superior àquela expansão de Edgeworth até $O(n^{-1/2})$. Exceto no ponto $y = 0$, a aproximação normal $\phi(y)$ para $\pi_{S_n^*}(y)$ não é satisfatória, como esperado, pois n é pequeno.

Exemplo 3.3 *Este exemplo (Barndorff-Nielsen e Cox, 1990, p.96) ilustra o desempenho da expansão de Edgeworth no contexto discreto. Seja S_n a soma de n variáveis aleatórias iid com distribuição de Poisson de média 1. Assim, S_n tem distribuição de Poisson de média n. Todos os cumulantes da distribuição de Poisson são iguais e, então, $\rho_3 = \rho_4 = 1$. A soma padronizada $S_n^* = (S_n - n)/\sqrt{n}$ tem função de distribuição aproximada, decorrente de (3.10), dada por*

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

No uso desta expansão para aproximar $P(S_n \leq r)$, pode-se adotar uma correção de continuidade como $y = (r - n + 0,5)/\sqrt{n}$ de modo que $P(S_n \leq r) = F_{S_n^*}(y)$.

A Tabela 3.2 compara a aproximação $\Phi(y)$ e as expansões de $F_{S_n^*}(y)$ até $O(n^{-1/2})$ e $O(n^{-1})$ com o valor exato de $P(S_n \leq r)$ quando $n = 8$. Ambas as expansões de Edgeworth aproximam melhor $P(S_n \leq r)$ do que a função de distribuição normal $\Phi(y)$.

Tabela 3.2: *Aproximações para a função de distribuição da Poisson de média* $n = 8$

			Expansões de Edgeworth	
r	Exato	Normal	até $O(n^{-1/2})$	até $O(n^{-1})$
2	0,0138	0,0259	0,0160	0,0148
4	0,0996	0,1079	0,1021	0,1011
6	0,3134	0,2981	0,3128	0,3141
8	0,5926	0,5702	0,5926	0,5919
10	0,8159	0,8116	0,8151	0,8146
12	0,9362	0,9442	0,9340	0,9374
14	0,9827	0,9892	0,9820	0,9824

Em inferência, o interesse principal reside em computar níveis de significância e, assim, a expansão (3.10) é mais útil do que a expansão para a função densidade (3.9). Frequentemente, em testes de hipóteses, trabalha-se com estatísticas padronizadas de média zero, variância um e cumulantes $\rho_j = \epsilon_j/\epsilon_2^{j/2}$ de ordens $O(n^{1-j/2})$ para $j \geq 3$. Neste caso, as probabilidades unilaterais envolvendo estatísticas padronizadas do tipo $P(Y_n \geq y)$ podem ser calculadas até $O(n^{-1})$ diretamente de (3.10) como

$$P(Y_n \geq y) = 1 - \Phi(y) + \phi(y)\left\{\frac{\rho_3}{6\sqrt{n}}H_2(y) + \frac{\rho_4}{24n}H_3(y) + \frac{\rho_3^2}{72n}H_5(y)\right\}, \qquad (3.11)$$

envolvendo um termo de ordem $O(n^{-1/2})$ e mais dois termos de ordem $O(n^{-1})$. Entretanto, as probabilidades bilaterais do tipo $P(|Y_n| \geq y)$ são obtidas (para $y > 0$) de (3.10) como

$$P(|Y_n| \geq y) = 2\{1 - \Phi(y)\} + 2\phi(y)\left\{\frac{\rho_4}{24n}H_3(y) + \frac{\rho_3^2}{72n}H_5(y)\right\},$$

envolvendo apenas correções de ordem $O(n^{-1})$. Neste caso, ocorre cancelamento das correções de ordem $O(n^{-1/2})$. Elas são iguais em magnitude, mas com sinais diferentes, e se cancelam quando as duas extremidades são combinadas.

Pode-se trabalhar com as expansões de Edgeworth (3.9) e (3.10) se as componentes Y_j são independentes mas não são necessariamente identicamente distribuídas. Tem-se $\kappa_r(S_n) = \Sigma\kappa_r(Y_j)$ e padroniza-se S_n na forma usual

$$S_n^* = \frac{S_n - \Sigma\kappa_1(Y_j)}{\Sigma\kappa_2(Y_j)}$$

e as expansões (3.9) e (3.10) continuarão valendo desde que as quantidades

$$\rho_3 = \frac{\Sigma\kappa_3(Y_j)}{\{\Sigma\kappa_2(Y_j)\}^{3/2}} \quad \text{e} \quad \rho_4 = \frac{\Sigma\kappa_4(Y_j)}{\{\Sigma\kappa_2(Y_j)\}^2}$$

sejam limitadas quando $n \longrightarrow \infty$.

3.4 Expansões de Cornish-Fisher

As expansões de Cornish-Fisher são usadas para determinar numericamente as distribuições de probabilidade de estatísticas quando suas distribuições exatas são difíceis de ser computadas. Suponha que uma variável aleatória contínua padronizada Y tem média zero, variância um e cumulantes ρ_j de ordens $O(n^{1-j/2})$ para $j \geq 3$. Neste caso, a expansão de Edgeworth para $P(Y \leq y)$ segue diretamente de (3.11). Suponha agora que y_α e u_α são definidos por $P(Y \leq y_\alpha) = \Phi(u_\alpha) = 1 - \alpha$. As *expansões de Cornish-Fisher* são duas expansões assintóticas relacionando os quantis y_α e u_α: uma expansão normalizadora que expressa u_α como função de y_α e sua expansão inversa dando y_α em termos de u_α.

A demonstração dessas expansões requer cálculos algébricos longos e apresenta-se aqui apenas a idéia da prova. Expandindo $\Phi(u_\alpha)$ vem

$$\Phi(u_\alpha) = \Phi\{y_\alpha + (u_\alpha - y_\alpha)\} = \Phi(y_\alpha) + \sum_{r=1}^{\infty} \frac{(u_\alpha - y_\alpha)^r}{r!} D^r \Phi(y_\alpha)$$

e, então,

$$\Phi(u_\alpha) = \Phi(y_\alpha) + \sum_{r=1}^{\infty} \frac{(u_\alpha - y_\alpha)^r}{r!} (-1)^{r-1} H_{r-1}(y_\alpha)\phi(y_\alpha). \tag{3.12}$$

Igualando $P(Y \leq y_\alpha)$ proveniente de (3.11) à equação (3.12), pode-se expressar u_α em função de y_α até $O(n^{-1})$ como

$$u_\alpha = p(y_\alpha) = y_\alpha - \frac{\rho_3}{6\sqrt{n}}(y_\alpha^2 - 1) + \frac{\rho_3^2}{36n}(4y_\alpha^2 - 7y_\alpha) - \frac{\rho_4}{24n}(y_\alpha^3 - 3y_\alpha)\,. \tag{3.13}$$

Então, qualquer probabilidade $P(Y \geq y_\alpha)$ até $O(n^{-1})$ é facilmente calculada como $1 - \Phi(u_\alpha)$, com o quantil u_α dado por (3.13). Este procedimento de cálculo é válido para qualquer estatística contínua padronizada que tenha terceiro e quarto cumulantes de

ordens $O(n^{-1/2})$ e $O(n^{-1})$, respectivamente, e os demais cumulantes de ordem $o(n^{-1})$.

O polinômio (3.13) de Cornish-Fisher representa a *transformação normalizadora* $p(Y)$ da variável Y até $O(n^{-1})$, isto é, $p(Y) \sim N(0,1) + O_p(n^{-3/2})$. Este polinômio é usado comumente para normalizar qualquer distribuição de probabilidade fazendo algum dos seus parâmetros tender para infinito, ou seja, substituindo-se n no resultado assintótico de grandes amostras ($n \longrightarrow \infty$) por algum parâmetro da distribuição de interesse que cresce indefinidamente. O exemplo a seguir ilustra isso.

Exemplo 3.4 *Considere o cálculo da expansão de Cornish-Fisher normalizadora da variável aleatória de Poisson* $Z \sim P(\lambda)$. *Padronizando-se esta variável* $Y = (Z - \lambda)/\sqrt{\lambda}$ *pode-se usar (3.13) com* $\lambda \longrightarrow \infty$ *ao invés de* $n \longrightarrow \infty$. *Observe-se que* $Y \xrightarrow{\mathcal{D}} N(0,1)$ *quando* $\lambda \longrightarrow \infty$. *Como* $\rho_3 = \lambda^{-1/2}$ *e* $\rho_4 = \lambda^{-1}$, *obtém-se*

$$p(Y) = Y - \frac{1}{6\sqrt{\lambda}}(Y^2 - 1) - \frac{1}{72\lambda}(3Y^3 - 8Y^2 + 5Y) \ .$$

Assim, a variável transformada $p(Y)$ *acima tem distribuição* $N(0,1)$ *com erro* $O(\lambda^{-3/2})$. *A expansão formal de Edgeworth para a distribuição de* $P(Z \leq z)$ *segue de (3.10) como*

$$P(Z \leq z) = \Phi(y) - \phi(y)\left\{\frac{y^2 - 1}{6\sqrt{\lambda}} + \frac{y^5 - 7y^3 + 3y}{72\lambda}\right\} + O(\lambda^{-3/2}),$$

sendo $y = (z - \lambda + 0{,}5)/\sqrt{\lambda}$ *com a correção 0,5 de continuidade.*

O objetivo da expansão inversa de Cornish-Fisher é expressar os quantis y_α de Y como função dos correspondentes quantis u_α da distribuição normal reduzida. A inversão da expansão (3.13) para calcular y_α em termos do quantil u_α da normal reduzida é feita através da fórmula geral de inversão de Lagrange. Então, $y_\alpha = u_\alpha + g(y_\alpha)$ pode ser expandido em termos de u_α como

$$y_\alpha - u_\alpha = g(u_\alpha) + \frac{Dg^2(u_\alpha)}{2!} + \frac{D^2 g^3(u_\alpha)}{3!} + \cdots \tag{3.14}$$

Identificando o polinômio $g(y_\alpha) = y_\alpha - p(y_\alpha)$ em (3.13), substituindo em (3.14) e calculando as potências de $g(u_\alpha)$ e suas derivadas, obtém-se y_α em função de u_α até $O(n^{-1})$

como

$$y_\alpha = u_\alpha + \frac{\rho_3}{6\sqrt{n}}(u_\alpha^2 - 1) - \frac{\rho_3^2}{36n}(2u_\alpha^3 - 5u_\alpha) + \frac{\rho_4}{24n}(u_\alpha^3 - 3u_\alpha). \qquad (3.15)$$

A importância da inversão de Cornish-Fisher (3.15) na inferência é possibilitar o cálculo dos quantis de estatísticas em termos dos quantis correspondentes da distribuição normal reduzida, conforme ilustra o exemplo abaixo.

Exemplo 3.5 *Suponha que* $Z \sim \chi_n^2$ *e seja* $Y = (Z - n)/\sqrt{2n}$ *a variável aleatória qui-quadrado padronizada, cujos terceiro e quarto cumulantes são* $\rho_3 = 2\sqrt{2}$ *e* $\rho_4 = 12$. *Logo,* $P(Z \leq z_\alpha) = P(Y \leq (z_\alpha - n)/\sqrt{2n})$ *e, portanto, juntando os dois termos de ordem* n^{-1} *em (3.15) vem*

$$z_\alpha = n + \sqrt{2n}\left\{u_\alpha + \frac{\sqrt{2}}{3\sqrt{n}}(u_\alpha^2 - 1) + \frac{1}{18n}(u_\alpha^3 - 7u_\alpha)\right\} .$$

A Tabela 3.3 (Barndorff-Nielsen e Cox, 1990, p.119) mostra a adequação das aproximações para z_α *provenientes da equação acima usando apenas o termo de ordem* $O(1)(u_\alpha)$ *e aquelas incluindo os termos* $O(n^{-1/2})$ *e* $O(n^{-1})$. *Observa-se desta tabela que a correção* $O(n^{-1/2})$ *já melhora substancialmente a aproximação normal, sendo que esta aproximação é ruim mesmo para* $n = 100$, *ao nível de significância de 1%.*

Tabela 3.3: *Comparação das expansões de Cornish-Fisher para os quantis da* χ_n^2

			Expansões até		
α	n	Exato	$O(1)$	$O(n^{-1/2})$	$O(n^{-1})$
	5	15,09	12,36	15,20	15,07
	10	23,21	20,40	23,34	23,25
0,01	50	76,15	73,26	76,20	76,16
	100	135,81	132,90	135,84	135,81
	5	9,24	9,65	9,48	9,24
	10	15,99	15,73	16,16	15,99
0,10	50	63,17	62,82	63,24	63,16
	100	118,50	118,12	118,55	118,50

3.5 Expansões Ponto de Sela

As expansões ponto de sela são muito importantes na teoria assintótica para aproximar com grande precisão as funções densidade e de distribuição, sendo facilmente deduzidas da função geratriz de cumulantes correspondente.

Sejam $Y_1, \ldots, Y_n$ variáveis aleatórias contínuas *iid* com função densidade $f(y)$ e funções geratrizes de momentos e cumulantes $M(t)$ e $K(t)$, respectivamente. Define-se a família exponencial conjugada de $f(y)$, indexada por um parâmetro λ, por

$$f(y;\lambda) = \exp\{\lambda y - K(\lambda)\} f(y) \; . \tag{3.16}$$

A família exponencial (3.16) 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 $M(t)$ de Y. A função geratriz de cumulantes $K(t;\lambda)$ correspondente a (3.16) é expressa em termos daquela $K(t)$ de Y por $K(t;\lambda) = K(t+\lambda) - K(\lambda)$.

Sejam $f_{S_n}(s;\lambda)$ e $K_{S_n}(t;\lambda)$ as funções densidade e geratriz de cumulantes de S_n relativas à família (3.16). Tem-se $K_{S_n}(t;\lambda) = nK(t+\lambda) - nK(\lambda)$ e, por inversão, vem

$$f_{S_n}(s;\lambda) = \exp\{s\lambda - nK(\lambda)\} f_{S_n}(s) \tag{3.17}$$

sendo $f_{S_n}(s) = f_{S_n}(s;0)$.

As funções densidade de S_n e S_n^* correspondentes à família (3.16) estão relacionadas por

$$f_{S_n}(s;\lambda) = f_{S_n^*}(y;\lambda) \frac{1}{\sqrt{nK''(\lambda)}} \; , \tag{3.18}$$

onde $y = \{s - nK'(\lambda)\}/\sqrt{nK''(\lambda)}$. Aproxima-se $f_{S_n^*}(y;\lambda)$ pela expansão de Edgeworth (3.9) escolhendo convenientemente $y = 0$ para anular o termo $O(n^{-1/2})$. Esta escolha equivale a considerar a distribuição em (3.16) definida por $\hat{\lambda}$ que satisfaz a equação $K'(\hat{\lambda}) = s/n$. Pode-se interpretar $\hat{\lambda}$ como a EMV de λ baseada numa única observação s de (3.17). Logo, $f_{S_n}(s;\hat{\lambda}) = f_{S_n^*}(0;\hat{\lambda})\{nK''(\hat{\lambda})\}^{-1/2}$. Agora, $f_{S_n^*}(0;\hat{\lambda})$ segue de (3.9),

observando que os cumulantes referentes a (3.16) são n vezes as derivadas de $K(\lambda)$

$$f_{S_n^*}(0;\hat{\lambda}) = \frac{1}{\sqrt{2\pi}}\{1 + M(\hat{\lambda})\} + O(n^{-2}), \tag{3.19}$$

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

$$M(\lambda) = \frac{3\rho_4(\lambda) - 5\rho_3(\lambda)^2}{24n}, \tag{3.20}$$

sendo $\rho_j(\lambda) = K^{(j)}(\lambda)/K^{(2)}(\lambda)^{j/2}$ para $j = 3$ e 4 e $K^{(j)}(\lambda) = d^jK(\lambda)/d\lambda^j$. Assim, $\rho_3(\lambda)$ e $\rho_4(\lambda)$ são os cumulantes padronizados que medem a assimetria e a curtose da distribuição (3.16). O erro em (3.19) é $O(n^{-2})$, pois o polinômio correspondente a $O(n^{-3/2})$ é de ordem ímpar e se anula em zero.

Fazendo $\lambda = \hat{\lambda}$ em (3.17), explicitando $f_{S_n}(s)$ e usando (3.18) e (3.19) vem

$$f_{S_n}(s) = \frac{\exp\{nK(\hat{\lambda}) - s\hat{\lambda}\}}{\sqrt{2n\pi K^{(2)}(\hat{\lambda})}}\{1 + M(\hat{\lambda}) + O(n^{-2})\}\,. \tag{3.21}$$

A fórmula (3.21) para aproximar a função densidade de S_n é denominada *expansão ponto de sela da soma* e produz aproximações precisas para funções densidades baseadas nas suas funções geratrizes de cumulantes. A terminologia é proveniente de uma dedução alternativa através da integral de contorno que inverte na função geratriz de momentos de S_n (Daniels, 1954). Observe-se que o termo principal de (3.21) só depende da função geratriz de cumulantes $K(t)$ de Y. Esta fórmula é bem diferente da expansão de Edgeworth (3.9). Primeiro, para usar (3.21) é necessário calcular, além de $\hat{\lambda}$, a função geratriz de cumulantes $K(t)$ de Y e não somente os seus 4 primeiros cumulantes. Entretanto, nas aplicações isso não apresenta grandes dificuldades. O termo principal em (3.21) não é a função densidade da distribuição normal $N(0,1)$ e, embora seja sempre positivo, nem sempre integra um. Entretanto, este termo pode ser normalizado. A expansão (3.21) é dada em potências de n^{-1}, enquanto a expansão de Edgeworth é dada em potências de $n^{-1/2}$. Uma desvantagem de (3.21) é que nem sempre é fácil integrar o seu lado direito para obter uma aproximação para a função de distribuição de S_n.

Verifica-se de imediato que a expansão ponto de sela para S_n^* num ponto qualquer w segue expressão idêntica à (3.21) com $nK^{(1)}(0) + \sqrt{n}wK^{(2)}(0)$ no lugar de s e o radicando

sendo substituido por $2\pi K^{(2)}(\hat{\lambda})/K^{(2)}(0)$. Esta expansão constitui, em geral, uma melhor aproximação para a função densidade exata de S_n^* do que (3.9), pois o erro é $O(n^{-2})$ ao invés de $O(n^{-3/2})$. Entretanto, na expansão ponto de sela, o erro é multiplicativo, enquanto na de Edgeworth é aditivo. A fórmula (3.21) é satisfeita mesmo para regiões de grandes desvios da forma $|s - nE(Y)| \leq b_n$, para b_n fixado, e em certos casos, mesmo para todos os valores de s (Jensen, 1988). Na Seção 5.4 apresenta-se uma aproximação para a função densidade da EMV baseada em (3.21).

A expansão para a função densidade da média amostral $\overline{Y}_n = S_n/n$ segue diretamente de (3.21) como

$$f_{\overline{Y}_n}(y) = \left\{\frac{n}{2\pi K^{(2)}(\hat{\lambda})}\right\}^{1/2} \exp[n\{K(\hat{\lambda}) - \hat{\lambda}y\}]\{1 + M(\hat{\lambda}) + O(n^{-2})\}, \tag{3.22}$$

onde $M(\lambda)$ é obtido de (3.20). O termo principal em (3.22) é denominado *aproximação ponto de sela para* $f_{\overline{Y}_n}(y)$. Assim, basta conhecer a função geratriz de cumulantes $K(t)$ comum de n variáveis aleatórias *iid* para se obter a aproximação ponto de sela da função densidade da média amostral dessas variáveis.

Exemplo 3.6 *Sejam* $Y_1, \ldots, Y_n$ *variáveis aleatórias iid com distribuição* $N(\mu, \sigma^2)$. *A função geratriz de cumulantes é* $K(\lambda) = \lambda\mu + \lambda^2\sigma^2/2$ *e a EMV* $\hat{\lambda}$ *é obtida de* $\mu + \hat{\lambda}\sigma^2 = s/n$. *Tem-se* $K^{(2)}(\hat{\lambda}) = \sigma^2$ *e* $K^{(3)}(\lambda) = K^{(4)}(\lambda) = 0$ *implicando* $M(\lambda) = 0$. *Logo, obtém-se de (3.21)*

$$f_{S_n}(s) = \frac{\exp\{-(s - n\mu)^2/(2n\sigma^2)\}}{\sqrt{2n\pi\sigma^2}}\{1 + O(n^{-2})\}.$$

O termo principal da expressão acima é a função densidade da distribuição $N(n\mu, n\sigma^2)$ *de* S_n. *Neste caso, a expansão ponto de sela reproduz a função densidade exata de* S_n.

Exemplo 3.7 *Considere a situação do exemplo 3.2 na qual* $Y_1, \ldots, Y_n$ *têm distribuição exponencial de média 1 e, então,* S_n *tem função densidade* $\pi_{S_n}(s) = s^{n-1}e^{-s}/(n-1)!$. *Assim,* $M(\lambda) = (1-\lambda)^{-1}$ *e* $K(\lambda) = -\log(1-\lambda)$. *A EMV* $\hat{\lambda}$ *é* $\hat{\lambda} = 1 - n/s$, $K(\hat{\lambda}) = \log(s/n)$ *e* $K^{(2)}(\hat{\lambda}) = s^2/n^2$. *O termo* $M(\hat{\lambda})$ *decorre de (3.20) como* $M(\hat{\lambda}) = -1/12n$. *Logo, a expansão ponto de sela (3.21) implica*

$$f_{S_n}(s) = \frac{s^{n-1}e^{-s}}{\sqrt{2\pi}e^{-n}n^{n-1/2}}\left\{1 - \frac{1}{12n} + O(n^{-2})\right\}.$$

A expansão acima está de acordo com a função densidade exata $\pi_{S_n}(s)$ podendo ser diretamente obtida a partir desta usando a aproximação de Stirling $(2\pi)^{1/2}e^{-n}n^{n-1/2}\{1+\frac{1}{12n}+O(n^{-2})\}$ para a função gama $\Gamma(n)=(n-1)!$.

Usualmente, o interesse maior em inferência reside em obter aproximações precisas para probabilidades do tipo $P(S_n \geq s)$ (ou $P(\overline{Y}_n \geq y)$) de uma amostra *iid* de n observações. A expansão de Edgeworth (3.10) pode ser usada com este objetivo, mas o erro da aproximação pode se tornar grande nas extremidades da distribuição de S_n (ou $\overline{Y}_n$). Uma maneira óbvia de aproximar $P(S_n \geq s)$ é integrar numericamente a aproximação ponto de sela representada pelo termo principal em (3.21), preservando as propriedades excelentes deste termo, ou seja, calcular

$$P(S_n \leq s) = \int_{-\infty}^{s} \frac{e^{\{nK(\hat{\lambda})-x\hat{\lambda}\}}}{\sqrt{2n\pi K^{(2)}(\hat{\lambda})}}dx \ .$$

O cálculo da integral acima é complicado e o leitor poderá consultar Daniels (1987), DiCiccio, Field e Fraser (1990), Barndorff-Nielsen e Cox (1990, Seção 4.3) e Hinkley, Reid e Snell (1991, Seção 12.4).

Pode-se demonstrar, com extensa álgebra, que a expansão de $P(S_n \geq s)$ até termos de ordem $O(n^{-1})$ quando $s > nE(Y)$, isto é, quando $\hat{\lambda} > 0$, é dada por (Daniels, 1987)

$$\begin{aligned} P(S_n \geq s) = \exp(n\hat{K}-s\hat{\lambda}+\hat{v}^2/2)\Bigg[\{1-\Phi(\hat{v})\}\left\{1-\frac{\hat{\rho}_3\hat{v}^3}{6\sqrt{n}}+\frac{1}{n}\left(\frac{\hat{\rho}_4\hat{v}^4}{24}+\frac{\hat{\rho}_3^2\hat{v}^6}{72}\right)\right\} \\ +\phi(\hat{v})\left\{\frac{\hat{\rho}_3(\hat{v}^2-1)}{6\sqrt{n}}-\frac{1}{n}\left(\frac{\hat{\rho}_4(\hat{v}^3-\hat{v})}{24}+\frac{\hat{\rho}_3^2(\hat{v}^5-\hat{v}^3+3\hat{v})}{72}\right)\right\}\Bigg], \end{aligned} \tag{3.23}$$

onde $\hat{\rho}_3 = \rho_3(\hat{\lambda})$, $\hat{\rho}_4 = \rho_4(\hat{\lambda})$, $\hat{K} = K(\hat{\lambda})$ e $\hat{v} = \hat{\lambda}\{nK^{(2)}(\hat{\lambda})\}^{1/2}$. A aproximação obtida de (3.23) com apenas os termos de ordem $O(\sqrt{n})$ fornece, em geral, bons resultados.

No caso de $s < nE(Y)$, ou seja, $\hat{\lambda} < 0$, pode-se obter $P(S_n \geq s)$ até $O(n^{-1/2})$ como

$$\begin{aligned} P(S_n \geq s) = H(-\hat{v}) + \exp(n\hat{K}-\hat{\lambda}s+\hat{v}^2/2)\times \\ \left[\{H(\hat{v})-\Phi(\hat{v})\}\left(1-\frac{\hat{\rho}_3\hat{v}^3}{6\sqrt{n}}\right)+\phi(\hat{v})\frac{\hat{\rho}_3(\hat{v}^2-1)}{6\sqrt{n}}\right], \end{aligned} \tag{3.24}$$

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

As equações (3.23) e (3.24) dependem do sinal de $\hat{\lambda}$, sendo (3.24) correta apenas até $O(n^{-1/2})$. Uma forma alternativa simples de obter $P(S_n \leq s)$ até $O(n^{-1})$, válida sobre todo o intervalo de variação de s, é devida a Lugannani e Rice (1980), que deduziram a seguinte fórmula:

$$P(S_n \leq s) = \Phi(\hat{r}) + \left(\frac{1}{\hat{r}} - \frac{1}{\hat{v}}\right)\phi(\hat{r}), \tag{3.25}$$

onde $\hat{r} = \text{sinal}(\hat{\lambda})[2n\{\hat{\lambda}K'(\hat{\lambda}) - K(\hat{\lambda})\}]^{1/2}$, cujo erro é $o(n^{-1})$ uniformemente em s.

As quantidades $\hat{r}$ e $\hat{v}$ podem ser interpretadas como a razão de verossimilhança sinalizada e a estatística escore (vide Seção 4.3), respectivamente, para testar $\lambda = 0$ no modelo exponencial (3.17) determinado por S_n.

A aproximação (3.25) é boa em quase todo o intervalo de variação de s, exceto próximo ao ponto $s = E(S_n)$ ou $r = 0$, onde deve ser substituída pelo seu limite, quando $r \longrightarrow 0$, dado por

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

Os exemplos 3.8 e 3.9 e as Tabelas 3.4 e 3.5 correspondentes ilustram para as distribuições exponencial e uniforme, respectivamente, a adequação das aproximações para $P(S_n \geq s)$ decorrentes de (3.23) incluindo os termos de ordens $O(n^{-1/2})$ e $O(n^{-1})$ e aquela aproximação dada por (3.25), onde estão expressos também os valores exatos de $P(S_n \geq s)$ para comparação.

Exemplo 3.8 *Suponha a distribuição exponencial de média um e função densidade* $f(y) = e^{-y}(y > 0)$. *Tem-se* $K(\lambda) = -\log(1-\lambda)$. *A Tabela 3.4 compara as três aproximações decorrentes de (3.23) e (3.25) e o valor exato de* $P(S_n \geq s)$ *para* $n = 1, 5$ *e* 10 *e diversos valores de* s. *Observe-se que (3.25) fornece resultados excelentes mesmo para* $n = 1$.

Exemplo 3.9 *Considere a distribuição uniforme com função densidade* $f(y) = \frac{1}{2}(-1 \leq y \leq 1)$ *e* $K(\lambda) = \log\{\text{senh}(\lambda)/\lambda\}$. *A Tabela 3.5 compara as três aproximações decorrentes de (3.23) e (3.25) e o valor exato de* $P(S_n \geq s)$ *para* $n = 1, 3$ *e* 10 *e diversos valores de* s. *Para* $n = 10$, *as aproximações (3.23) até* $O(n^{-1})$ *e (3.25) praticamente se igualam aos valores exatos.*

Tabela 3.4: *Comparação das aproximações ponto de sela para* $P(S_n \geq s)$ *na distribuição exponencial*

			Aproximação (3.23)		
n	s	Exato	até $O(n^{-1/2})$	até $O(n^{-1})$	(3.25)
	0,5	0,6065	0,6176	0,6077	0,6043
1	1,0	0,3679	0,3670	0,3670	0,3670
	3,0	0,0498	0,0482	0,0510	0,0500
	7,0	0,00091	0,00095	0,00091	0,00093
5	1,0	0,99634	0,99638	0,99635	0,99633
	3,0	0,8153	0,8172	0,8156	0,8152
	5,0	0,4405	0,4405	0,4405	0,4405
	10,0	0,0293	0,0291	0,0293	0,0293
	20,0	0,0000169	0,0000171	0,0000169	0,0000170
	5,0	0,9682	0,9683	0,9682	0,9682
10	10,0	0,4579	0,4579	0,4579	0,4579
	15,0	0,0699	0,0695	0,0699	0,0699
	20,0	0,00500	0,00499	0,00500	0,00500

Tabela 3.5: *Comparação das aproximações ponto de sela para* $P(S_n \geq s)$ *na distribuição uniforme*

			Aproximação (3.23)		
n	s	Exato	até $O(n^{-1/2})$	até $O(n^{-1})$	(3.25)
	0,2	0,4	0,3897	0,3841	0,3838
1	0,4	0,3	0,2831	0,2767	0,2750
	0,6	0,2	0,1855	0,1830	0,1791
	0,8	0,1	0,0945	0,0974	0,0948
3	0,5	0,3177	0,3193	0,3168	0,3168
	1,0	0,1667	0,1699	0,1676	0,1673
	1,5	0,0703	0,0710	0,0699	0,0695
	2,5	0,00260	0,00255	0,00258	0,00254
	1,0	0,2945	0,2953	0,2945	0,2945
10	3,0	0,0505	0,0508	0,0505	0,0504
	5,0	0,00247	0,00249	0,00247	0,00246
	7,0	0,0000159	0,0000160	0,0000159	0,0000159

As expansões ponto de sela (3.23) - (3.25) só são válidas para variáveis aleatórias contínuas. No caso discreto, elas podem ser adaptadas com correções de continuidade. A expansão para $P(S_n \geq s)$ até $O(n^{-1/2})$ correspondente a (3.23) quando $s > nE(Y)$, válida para distribuições discretas, tem a forma (Daniels, 1987)

$$P(S_n \geq s) = \exp\{(\hat{r}^2 + \hat{v}^2)/2\}\{\hat{\lambda}/(1 - e^{-\hat{\lambda}})\}\times$$

$$\left[(1 - \Phi(\hat{v}))\left\{1 - \frac{\hat{\rho}_3\hat{v}^3}{6\sqrt{n}} - \frac{\hat{v}}{\sqrt{n\hat{K}''}}(\hat{\lambda}^{-1} - (e^{\hat{\lambda}} - 1)^{-1})\right\}\right. \tag{3.26}$$

$$\left. +\phi(\hat{v})\left\{\frac{\hat{\rho}_3(\hat{v}^2 - 1)}{6\sqrt{n}} + \frac{1}{\sqrt{n\hat{K}''}}(\hat{\lambda}^{-1} - (e^{\hat{\lambda}} - 1)^{-1})\right\}\right],$$

com todas as quantidades já definidas anteriormente.

A fórmula de Lugannani e Rice (3.25) pode ser aplicada no contexto discreto com as correções de continuidade para $\hat{\lambda}$ e $\hat{v}$ dadas por

$$nK'(\hat{\lambda}) = s - 0,5 \quad \text{e} \quad \hat{v} = (1 - e^{-\hat{\lambda}})\{nK^{(2)}(\hat{\lambda})\}^{1/2}.$$

Exemplo 3.10 *Ilustra-se na Tabela 3.6 o desempenho das equações (3.25) (com as correções de continuidade acima) e (3.26) para aproximar $P(S_n \geq s)$ no caso da distribuição de Poisson com média μ, onde $K(\lambda) = \mu(e^{\lambda} - 1)$, supondo $\mu = 0,2$, $n = 1$ e $\mu = 1$, $n = 1,5$ e 10, e considerando vários valores de s. A Tabela 3.6 mostra que o desempenho da fórmula (3.25) é excelente mesmo no caso discreto com $n = 1$.*

Outros exemplos numéricos apresentados por Daniels (1983, 1987) e Davison e Hinkley (1988) sinalizam para o uso em inferência da fórmula (3.25) no cálculo aproximado de probabilidades não somente associadas com somas e médias de variáveis aleatórias mas com inúmeras distribuições contínuas e discretas.

Tabela 3.6: *Comparação das aproximações ponto de sela para* $P(S_n \geq s)$ *na distribuição de Poisson*

$\mu = 0,2, \quad n = 1$

s	Exato	(3.25)	(3.26)
1	0,1813	0,1840	0,1759
2	0,0175	0,0177	0,0171
3	0,00115	0,00116	0,00112
4	0,0000568	0,0000572	0,0000563

$\mu = 1, \quad n = 1$

s	Exato	(3.25)	(3.26)
1	0,6321	0,6330	0,6330
3	0,0803	0,0804	0,0790
7	0,0000832	0,0000834	0,0000825
9	0,00000113	0,00000113	0,00000115

$\mu = 1, \quad n = 5$

s	Exato	(3.25)	(3.26)
1	0,99326	0,99319	0,99356
3	0,8753	0,8752	0,8765
5	0,5595	0,5595	0,5595
15	0,000226	0,000226	0,000225

$\mu = 1, \quad n = 10$

s	Exato	(3.25)	(3.26)
1	0,9999546	0,9999536	0,9999567
5	0,9707	0,9710	0,9710
10	0,5421	0,5421	0,5421
20	0,00345	0,00345	0,00344

3.6 Expansões de Laplace

As expansões assintóticas para muitas integrais usadas em Estatística, incluindo aproximações para funções de distribuição tais como função gama e funções de Bessel, podem ser deduzidas por uma técnica denominada de *método de Laplace*. O interesse inicial é obter a expansão da *transformada de Laplace* $\mathcal{L}(z) = \int_0^\infty e^{-zy} f(y)dy$ para z grande. A função geratriz de momentos $M(t)$ da distribuição com função densidade $f(y)$ sobre os reais não-negativos é dada por $M(t) = \mathcal{L}(-t)$. Para funções $f(y)$ bem comportadas, a forma de $\mathcal{L}(z)$ para z grande é determinada pelos valores de $f(y)$ próximos a $y = 0$. Expandindo $f(y)$ em série de Taylor vem

$$f(y) = \sum_r f^{(r)}(0)\frac{y^r}{r!}$$

e, então,

$$\mathcal{L}(z) = \int_0^\infty e^{-zy}\left(\sum_r f^{(r)}(0)\frac{y^r}{r!}\right)dy$$

ou

$$\mathcal{L}(z) = \sum_r \frac{f^{(r)}(0)}{r!}\int_0^\infty e^{-zy}y^r dy\ .$$

Como a integral acima iguala $r!/z^{r+1}$, obtém-se

$$\mathcal{L}(z) = \sum_r \frac{f^{(r)}(0)}{z^{r+1}} = \frac{f(0)}{z} + \frac{f'(0)}{z^2} + \cdots \tag{3.27}$$

Exemplo 3.11 *Considere a determinação da expansão da integral da normal* $\Phi(z) = 1 - \int_z^\infty \phi(y)dy$ *para* z *grande. Por simples mudança de variáveis vem*

$$\Phi(z) = 1 - \phi(z) \int_0^\infty e^{-zt} e^{-t^2/2} dt \ .$$

Fazendo $f(t) = e^{-t^2/2}$ *e calculando a expansão da integral acima usando (3.27), tem-se*

$$\Phi(z) = 1 - \frac{\phi(z)}{z} \left\{ 1 - \frac{1}{z^2} + \frac{3}{z^4} - \frac{7}{2z^6} + \ldots \right\} \ . \tag{3.28}$$

Para z *fixado, o erro cometido no trunamento de (3.28) é menor do que o primeiro termo omitido, embora a série infinita seja divergente. Claramente, fixado o número de termos em (3.28), a aproximação melhora quando* z *cresce.*

Considere agora que a integral a ser avaliada para $z \longrightarrow \infty$ tem a forma

$$w(z) = \int_a^b e^{-zr(y)} f(y) dy \ . \tag{3.29}$$

O cálculo da expansão da integral (3.29) para z grande é útil para aproximar várias integrais de interesse na Estatística. A contribuição principal para $w(z)$ quando z é grande vem dos valores de y próximos ao mínimo de $r(y)$ que pode ocorrer em a ou b ou no interior do intervalo (a, b). Suponha, inicialmente, que $r(y)$ é minimizada em $\tilde{y} \in (a, b)$ e que $r'(\tilde{y}) = 0$, $r''(\tilde{y}) > 0$ e $f(\tilde{y}) \neq 0$. Tem-se,

$$w(z) = \int_a^b \exp\{-z\tilde{r} - z(y - \tilde{y})^2 \tilde{r}''/2 - \cdots\} f(y) dy$$

com a convenção $\tilde{r} = r(\tilde{y})$, $\tilde{r}'' = r''(\tilde{y})$, $\tilde{f} = f(\tilde{y})$, etc. Ainda,

$$w(z) = e^{-z\tilde{r}} \sqrt{\frac{2\pi}{z\tilde{r}''}} \int_{-\infty}^{+\infty} \{\tilde{f} + (y - \tilde{y})\tilde{f}' + \ldots\} \phi\left(y - \tilde{y};\ \frac{1}{y\tilde{r}''}\right) dy \ ,$$

onde $\phi(y-\mu;\sigma^2)$ representa a função densidade da distribuição normal $N(\mu,\sigma^2)$. Com alguma álgebra, demonstra-se (Barndorff-Nielsen e Cox, 1990, Seção 3.3) que $w(z)$ pode ser escrita até $O(z^{-1})$ como

$$w(z) = e^{-z\tilde{r}}\sqrt{\frac{2\pi}{z\tilde{r}''}}\left\{\tilde{f}+\frac{1}{z}\left(\frac{\tilde{f}''}{2\tilde{r}''}-\frac{\tilde{r}^{(3)}\tilde{f}'}{2\tilde{r}''^2}-\frac{\tilde{r}^{(4)}\tilde{f}}{8\tilde{r}''^2}+\frac{5(\tilde{r}^{(3)})^2\tilde{f}}{24\tilde{r}''^3}\right)+O(z^{-2})\right\}. \qquad (3.30)$$

No caso de $r(y)$ ser minimizada em $\tilde{y}=a$ (ou b) e $r'(\tilde{y})$ não sendo nulo, obtém-se

$$w(z) = e^{-z\tilde{r}}\left\{\frac{\tilde{f}}{z\tilde{r}'}+O(z^{-2})\right\}.$$

Outros refinamentos do método de Laplace incluindo a possibilidade de $r(y)$ depender fracamente de z são apresentados no livro de Barndorff-Nielsen e Cox (1990, Seção 3.3).

Exemplo 3.12 *Seja o cálculo da função gama* $\Gamma(z+1)=\int_0^\infty x^z e^{-x}dx$ *para* z *grande. Com a mudança de variável* $y=x/z$ *vem*

$$\Gamma(z+1) = z^{z+1}\int_0^\infty \exp(z\log y - zy)dy$$

que é exatamente da forma (3.29) com $f(y)=1$ *e* $r(y)=-\log y+y$. *Tem-se* $\tilde{y}=1$, $\tilde{r}=1$, $\tilde{r}'=0$, $\tilde{r}''=1$, $\tilde{r}^{(3)}=-2$ *e* $\tilde{r}^{(4)}=6$. *Substituindo esses valores em (3.30) vem*

$$\Gamma(z+1) = \sqrt{2\pi}z^{z+1/2}e^{-z}\left\{1+\frac{1}{12z}+O(z^{-2})\right\} \qquad (3.31)$$

que é a expansão de Stirling. A aproximação (3.31) é boa para $z \geq 1{,}5$.

3.7 Expansões Assintóticas para Variáveis Aleatórias

Algumas vezes é mais fácil aproximar as variáveis aleatórias de interesse diretamente do que obter aproximações através de suas funções de distribuição. Sejam X_0, X_1 e X_2 variáveis aleatórias contínuas com funções densidade marginais não dependentes de n e tendo suporte em $\mathbb{R}$. Considere a seqüência de variáveis aleatórias $\{Y_n\}$ definida quando

$n \longrightarrow \infty$ por

$$Y_n = X_0 + n^{-1/2}X_1 + n^{-1}X_2 + O_p(n^{-3/2}) \; . \tag{3.32}$$

Uma expansão como (3.32) é denominada *expansão estocástica assintótica.* Vários exemplos de expansões do tipo (3.32) aparecem na literatura estatística. O objetivo principal é calcular a função de distribuição $F_n(y) = P(Y_n \leq y)$ de Y_n até ordem n^{-1} em termos das funções de distribuição $F_0(y) = P(X_0 \leq y)$ e densidade $f_0(y) = \frac{dF_0(y)}{dy}$ de X_0 e de certos valores esperados de X_1 e X_2 condicionados a $X_0 = y$. Expansões estocásticas assintóticas e expansões assintóticas para funções de distribuição são equivalentes supondo a validade de certas condições de regularidade, conforme determina o seguinte teorema de Cox e Reid (1987):

Teorema de Cox e Reid

A função de distribuição $F_n(y)$ da variável aleatória Y_n definida por (3.32), supondo certas condições gerais, é dada até $O(n^{-1})$ por

$$F_n(y) = F_0(y)\{1 + n^{-1/2}a_1(y) + n^{-1}a_2(y)\}, \tag{3.33}$$

onde as funções $a_1(y)$ e $a_2(y)$ são determinadas a partir das equações

$$F_0(y)a_1(y) = -E(X_1|X_0 = y)\; f_0(y), \tag{3.34}$$

$$F_0(y)a_2(y) = -E(X_2|X_0 = y)\; f_0(y) + \frac{1}{2}\frac{\partial}{\partial y}\{E(X_1^2|X_0 = y)f_0(y)\} \; . \tag{3.35}$$

A recíproca do teorema acima é também verdadeira e pode-se construir Y_n em (3.32) a partir de (3.33) definindo convenientemente X_0, X_1 e X_2 para satisfazer (3.34) - (3.35). A equivalência entre as expansões (3.32) e (3.33) é importante na teoria assintótica, conforme será mostrado nos dois exemplos seguintes e na Seção 5.7.

Exemplo 3.13 *Como ilustração da aplicabilidade do teorema de Cox e Reid mostra-se como obter a expansão de Edgeworth (3.11) para a função de distribuição de Y_n a partir da expansão de Cornish-Fisher (3.15) de Y_n. Assim, a expansão estocástica assintótica*

até $O(n^{-1})$ *dada em (3.15) é*

$$Y_n = U + \frac{\rho_3}{6\sqrt{n}}(U^2-1) - \frac{\rho_3^2}{36n}(2U^3-5U) + \frac{\rho_4}{24n}(U^3-3U)$$

com $U \sim N(0,1)$. *Identificando* $X_0 = U$, $f_0(y) = \phi(y)$, $X_1 = \rho_3(U^2-1)/6$ *e* $X_2 = -\rho_3^2(2U^3-5U)/36 + \rho_4(U^3-3U)/24$ *vem*

$$E(X_1|U=y) = \rho_3(y^2-1)/6\ ,$$
$$E(X_2|U=y) = -\rho_3^2(2y^3-5y)/36 + \rho_4(y^3-3y)/24\ ,$$
$$E(X_1^2|U=y) = \rho_3^2(y^2-1)^2/36$$

e

$$\frac{\partial}{\partial y}\{\rho_3^2(y^2-1)^2\phi(y)/36\} = -\rho_3^2(y^5-6y^3+5y)\phi(y)/36\ .$$

Logo, de (3.34) e (3.35) obtém-se

$$F_0(y)a_1(y) = -\rho_3(y^2-1)\phi(y)/6$$

e

$$F_0(y)a_2(y) = \{-\rho_4(y^3-3y)/24 + \rho_3^2(2y^3-5y)/36\}\phi(y) - \rho_3^2(y^5-6y^3+5y)\phi(y)/72$$

ou

$$F_0(y)a_2(y) = -\rho_4 H_3(y)\phi(y)/24 - \rho_3^2 H_5(y)\phi(y)/72\ .$$

Finalmente, substituindo-se em (3.33) chega-se à expansão de Edgeworth (3.11).

Exemplo 3.14 *Suponha a variável aleatória qui-quadrado padronizada* $Y_n = (\chi_n^2 - n)/\sqrt{2n}$ *cujos terceiro e quarto cumulantes são* $\rho_3 = 2\sqrt{2}$ *e* $\rho_4 = 12$ *(vide exemplo 3.5). Mostra-se aqui como se obtém a inversão de Cornish-Fisher para* Y_n *a partir da expansão de Edgeworth para a sua função de distribuição e do teorema de Cox e Reid. A expansão para a função de distribuição de* Y_n *até* $O(n^{-1})$ *segue de (3.11) como*

$$F_n(y) = \Phi(y) - \phi(y)\left\{\frac{\sqrt{2}}{3\sqrt{n}}H_2(y) + \frac{1}{2n}H_3(y) + \frac{1}{9n}H_5(y)\right\}.$$

Define-se $X_0 = U \sim N(0,1)$ *e, então,* $f_0(y) = \phi(y)$. *Consideram-se* X_1 *e* X_2 *como funções dependentes apenas de* U, $X_1 = \rho_1(U)$ *e* $X_2 = \rho_2(U)$, *a serem determinadas. Comparando os termos de ordem* $O(n^{-1/2})$ *da expansão acima e de (3.33), obtém-se de (3.34)*

$$-\phi(y)\frac{\sqrt{2}}{3}H_2(y) = -E\{\rho_1(U)|U = y\}\phi(y) = -\rho_1(y)\phi(y).$$

Logo, $X_1 = \frac{\sqrt{2}}{3}(U^2-1)$. *Analogamente, comparando os termos de ordem* $O(n^{-1})$, *obtém-se de (3.35)*

$$\begin{aligned} -\phi(y)\left\{\frac{1}{2}H_3(y) + \frac{1}{9}H_5(y)\right\} &= -E\{X_2|U = y\}\phi(y) + \frac{1}{2}\frac{\partial}{\partial y}\left\{\frac{2}{9}(y^2-1)^2\phi(y)\right\} \\ &= -\rho_2(y)\phi(y) + \frac{1}{9}\{-(y^2-1)^2y + 4y(y^2-1)\}\phi(y). \end{aligned}$$

Assim,

$$\rho_2(y) = \frac{1}{2}H_3(y) + \frac{1}{9}H_5(y) + \frac{1}{9}\{-(y^2-1)^2y + 4y(y^2-1)\}$$

que pela substituição dos polinômios de Hermite reduz-se a $\rho_2(y) = \frac{1}{18}(y^3 - 7y)$. *Finalmente,* $X_2 = \frac{1}{18}(U^3 - 7U)$ *e a fórmula (3.32) do teorema de Cox e Reid implica*

$$Y = U + \frac{\sqrt{2}}{3\sqrt{n}}(U^2-1) + \frac{1}{18n}(U^3 - 7U)\ .$$

Este resultado é idêntico àquele obtido no exemplo 3.5 usando diretamente a fórmula da inversão de Cornish-Fisher.

3.8 Expansões por Métodos Diretos

Muitas expansões do tipo (3.1) podem ser deduzidas para funções densidade e de distribuição e para funções geratrizes de momentos e cumulantes através dos *métodos diretos*, que consistem em padronizar a variável aleatória de interesse e expandir as funções matemáticas que dependem de n. Algumas vezes é mais conveniente expandir as funções geratrizes de momentos ou cumulantes e depois inverter termo a termo para obter as expansões das funções de distribuição e densidade. A seguir, apresentam-se alguns exemplos de expansões deduzidas pelos métodos diretos.

Exemplo 3.15 *Seja a função densidade da distribuição t de Student com n graus de liberdade dada por*

$$g_n(y) = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi}\Gamma(\frac{n}{2})}(1+y^2/n)^{-(n+1)/2}, \quad y \in I\!R\,.$$

A variável aleatória t de Student tem média zero e variância diferente de um, mas pode-se obter a expansão de $\log g_n(y)$ *a partir das expansões calculadas diretamente*

$$\begin{aligned} -\log\left(1+\frac{y^2}{n}\right) &= -\frac{y^2}{n}+\frac{y^4}{2n^2}-\cdots+\frac{(-y^2)^j}{jn^j}+\cdots, \\ -\frac{1}{2}(n+1)\log\left(1+\frac{y^2}{n}\right) &= -\frac{y^2}{2}+\cdots-\frac{j(-y^2)^{j+1}+(j+1)(-y^2)^j}{2j(j+1)n^j}+\cdots \end{aligned}$$

e de

$$\log\left\{\frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})}\right\} = \frac{1}{2}\log\left(\frac{n}{2}\right)-\frac{1}{4n}+\frac{1}{24n^3}-\frac{1}{20n^5}+\cdots$$

obtida da expansão de Stirling para $\log\Gamma(n+1)$. *Assim,*

$$\begin{aligned} \log g_n(y) = \; & -\frac{1}{2}\log(2\pi)-\frac{y^2}{2}+\frac{1}{4n}(y^4-2y^2-1) \\ & -\frac{1}{12n^2}(2y^6-3y^4)+\frac{1}{24n^3}(3y^8-4y^6+1)+O(n^{-4})\,. \end{aligned}$$

Tomando a exponencial da expressão anterior, obtém-se

$$\begin{aligned} g_n(y) = \; & \phi(y)\Big\{1+\frac{1}{4n}(y^4-2y^2-1)+\frac{1}{96n^2}(3y^8 \\ & -16y^6-12y^5+18y^4+12y^2+12y^2+12y+3)+O(n^{-3})\Big\}\,. \end{aligned} \tag{3.36}$$

Da expansão (3.36) verifica-se facilmente que a distribuição t de Student tende para a distribuição normal reduzida quando $n \longrightarrow \infty$.

Exemplo 3.16 *A distribuição de Poisson* $P(\lambda)$ *pode ser considerada como o limite da distribuição binomial* $B(n,\theta)$ *fazendo* $n \longrightarrow \infty$, $\theta \longrightarrow 0$ *com* $n\theta = \lambda$ *fixado. O logaritmo da probabilidade* π_r *de r sucessos na distribuição binomial é dado por (r fixo)*

$$\begin{aligned}\log \pi_r &= \log\left(\frac{\lambda^r}{r!}\right) + (n-r)\log\left(1-\frac{\lambda}{n}\right) + \log\left(1-\frac{1}{n}\right) + \log\left\{1-\frac{(r-1)}{n}\right\} \\ &= \log\left(\frac{\lambda^r}{r!}\right) - \lambda + \frac{1}{n}\left(r\lambda - \frac{\lambda^2}{2} - \frac{r^2}{2} + \frac{r}{2}\right) + O(n^{-2}) .\end{aligned}$$

Então,

$$\pi_r = \frac{e^{-\lambda}\lambda^r}{r!}\left\{1 + \frac{1}{n}\left(r\lambda - \frac{\lambda^2}{2} - \frac{r^2}{2} + \frac{r}{2}\right) + O(n^{-2})\right\} . \tag{3.37}$$

A expansão (3.37) mostra que a probabilidade da distribuição binomial é aproximada por uma probabilidade associada à distribuição de Poisson, com erro $O(n^{-2})$.

3.9 Expansões de Funções Não-Lineares

Nas seções anteriores, a discussão se referia a somas (ou médias) de variáveis aleatórias *iid.* Discute-se aqui uma generalização útil nas aplicações da teoria assintótica referente a uma função não-linear de uma soma (ou média) de variáveis aleatórias independentes. Por exemplo, a EMV em muitos problemas é uma função não-linear da soma (ou média) das observações.

Seja T_n uma estatística qualquer tal que $T_n \xrightarrow{P} \theta$ e suponha que $\sqrt{n}(T_n - \theta)$ tem distribuição normal $N(0,1)$ assintótica. Admite-se que $\sqrt{n}(T_n - \theta)$ tem uma expansão de Edgeworth do tipo (3.33) calculada a partir dos quatro primeiros momentos de T_n. Neste caso, o teorema de Cox e Reid (Seção 3.7) garante que é possível encontrar, a partir das equações (3.34) - (3.35), as funções $\rho_1(\cdot)$ e $\rho_2(\cdot)$ de uma variável aleatória $X \sim N(0,1)$ tal que

$$\sqrt{n}(T_n - \theta) = X + \frac{\rho_1(X)}{\sqrt{n}} + \frac{\rho_2(X)}{n} + O_p(n^{-3/2}) .$$

Seja $g(t)$ uma função não-linear de t bem comportada. Deseja-se obter a expansão estocástica assintótica para $\sqrt{n}\{g(T_n) - g(\theta)\}$ e calcular a expansão de sua função de distribuição. Tem-se que

$$g(T_n) = g\left\{\theta + \frac{X}{\sqrt{n}} + \frac{\rho_1(X)}{n} + \frac{\rho_2(X)}{n\sqrt{n}} + O_p(n^{-2})\right\} .$$

Expandindo a equação anterior em série de Taylor vem

$$\begin{aligned}\sqrt{n}\{g(T_n)-g(\theta)\} &= Xg'(\theta)+\left\{\rho_1(X)g'(X)+\frac{1}{2}X^2g''(\theta)\right\}/\sqrt{n} \\ &+\left\{\rho_2(X)g'(\theta)+X\rho_1(X)g''(\theta)+\frac{1}{6}X^3g'''(\theta)\right\}/n+O_p(n^{-3/2})\ .\end{aligned} \tag{3.38}$$

A equação (3.38) representa uma expansão estocástica assintótica do tipo (3.32) com $X_0 = Xg'(\theta)$, $X_1 = \rho_1(X)g'(X)+\frac{1}{2}X^2g''(\theta)$ e $X_2 = \rho_2(X)g'(\theta)+X\rho_1(X)g''(\theta)+\frac{1}{6}X^3g'''(\theta)$ e, portanto, admite uma expansão de Edgeworth do tipo (3.33), cujas funções $a_1(\cdot)$ e $a_2(\cdot)$ podem ser deduzidas com algum algebrismo das equações (3.34) - (3.35). Resumindo, funções não-lineares de estatísticas que possuem expansões de Edgeworth admitem tais expansões que podem ser deduzidas do teorema de Cox e Reid.

3.10 Aproximação Normal para Algumas Variáveis Discretas

As aplicações das expansões de Edgeworth e ponto de sela para variáveis aleatórias discretas envolvem o uso das *correções de continuidade*, que representa um método simples de avaliar probabilidades quando uma distribuição discreta é aproximada por uma contínua. Em muitas aplicações, a distribuição contínua que serve como aproximação é a distribuição normal e o método consiste em aproximar uma probabilidade do tipo $P(Y = y)$ de uma distribuição discreta por um intervalo correspondente $P(y-0,5 \leq Y \leq y+0,5)$ da distribuição normal supondo que Y varia de um em um. Similarmente, uma probabilidade tal qual $P(Y \leq y)$ de uma distribuição discreta pode ser aproximada por $P(Y \leq y+0,5)$ da distribuição normal correspondente. O ajustamento de y pela adição e subtração de 0,5 é uma *correção de continuidade*. A correção objetiva transformar um ponto y de um conjunto discreto, num intervalo $[y-0,5, y+0,5]$ contínuo, de modo que o valor aproximado da probabilidade pontual $P(Y = y)$ seja obtido como uma área correspondente ao intervalo unitário centrado em y e abaixo da função densidade usada na aproximação contínua. As distribuições discretas mais comuns onde são aplicadas as correções de continuidade são: binomial, Poisson, binomial negativa e hipergeométrica. No que se segue a probabilidade $P = P(Y \leq k|\theta)$, onde θ representa parâmetros, é aproximada por

$\Phi(u)$, onde u é uma função simples de k e θ e $\Phi(\cdot)$ é a função de distribuição acumulada da distribuição $N(0,1)$.

Distribuição Binomial

Se $Y \sim B(n,p)$, então

$$P = P(Y \leq k) = \sum_{j=0}^{k} \binom{n}{j} p^j (1-p)^{n-j}$$

para $k = 0, \ldots, n$. Pode-se usar $P \doteq \Phi((k+0,5-np)/(np(1-p))^{1/2})$ quando $\min(p, 1-p) > 5/n$. Este resultado é válido assintoticamente quando $n \longrightarrow \infty$ e $k \longrightarrow \infty$, de modo que $(y-np)^3/\{np(1-p)\}^2 \longrightarrow 0$. O erro absoluto máximo desta aproximação é menor do que $0,140\sqrt{np(1-p)}$. Um resultado aproximado equivalente é

$$P \doteq \Phi\left(\frac{k-np}{\sqrt{np(1-p)}}\right) + \frac{1}{2\sqrt{np(1-p)}}\phi\left(\frac{k-np}{\sqrt{np(1-p)}}\right),$$

onde $\phi(\cdot)$ é a função densidade da distribuição $N(0,1)$. Melhores aproximações para P são obtidas das equações

$$P \doteq \Phi\left(2[\{(k+1)(1-p)\}^{1/2} - \{(n-k)p\}^{1/2}]\right) \tag{3.39}$$

ou

$$P \doteq \Phi\left(\{(4k+3)(1-p)\}^{1/2} - \{(4n-4k-1)p\}^{1/2}\right) . \tag{3.40}$$

Usa-se (3.39) quando $p \leq 0,05$ ou $p \geq 0,93$ e (3.40) se $0,05 < p < 0,93$. Uma aproximação mais precisa é dada por Pratt (1968): $P(Y \leq k) \doteq \Phi(u)$, onde

$$u = d\frac{\left\{1 + \frac{(1-p)}{np}g(k+0,5) + \frac{p}{n(1-p)}g(n-k-0,5)\right\}^{1/2}}{\left\{\left(n+\frac{1}{6}\right)p(1-p)\right\}^{1/2}} \tag{3.41}$$

com $d = k + \frac{2}{3} - (n+\frac{1}{3})p$ e $g(x) = (1 - x^2 + 2x\log x)(1-x)^{-2}$ sendo $g(1) = 0$.

A aproximação (3.41) tem erro de ordem de magnitude $\{np(1-p)\}^{-3/2}$ uniformemente em k. Nos casos triviais $k=0$ e $k=n-1$, onde $P=(1-p)^n$ e $P=1-p^n$, respectivamente, esta aproximação se deteriora.

As probabilidades individuais $P(Y=k)$ podem ser computadas por

$$P(Y=k)\doteq\frac{1}{\sqrt{2\pi np(1-p)}}\exp\left\{-\frac{(k-np)^2}{2np(1-p)}\right\}.$$

Distribuição de Poisson

Se $Y\sim P(\lambda)$, então

$$P=P(Y\leq k)=\sum_{j=0}^{k}\frac{e^{-\lambda}\lambda^j}{j!}.$$

A probabilidade P acima pode ser computada exatamente a partir da função de distribuição qui-quadrado usando

$$P=P(Y\leq k)=P(\chi^2_{2(k+1)}\geq 2\lambda).$$

A aproximação clássica para P é obtida do teorema central do limite como $\Phi((k+0,5-\lambda)\lambda^{-1/2})$. Mesmo para λ grande sua precisão não é boa: para $\lambda=30, k=17$, resulta em $0,0113$ enquanto o valor exato é $0,0073$. Uma aproximação mais precisa para P é $P=P(Y\leq k)\doteq 1-\Phi(w)$, onde $w=3[(\frac{\lambda}{k+1})^{1/3}-1+\frac{1}{9(k+1)}](k+1)^{1/2}$, sendo baseada na aproximação de Wilson-Hilferty para a distribuição qui-quadrado. Uma outra aproximação simples supõe que $2(\sqrt{Y}-\sqrt{\lambda})$ tem distribuição normal $N(0,1)$.

Aproximações aperfeiçoadas para $P=P(Y\leq k)$ podem ser obtidas de

$$P\doteq\Phi(2\{(k+1)^{1/2}-\lambda^{1/2}\}) \tag{3.42}$$

e

$$P\doteq\Phi((4k+3)^{1/2}-2\lambda^{1/2}). \tag{3.43}$$

A aproximação (3.42) é bastante adequada próximo aos níveis de significância usuais enquanto (3.43) funciona melhor se $0,05<P<0,93$. Uma aproximação alternativa para

P pode ser deduzida da expansão (3.13) de Cornish-Fisher. Assim, $P = P(Y \leq k) \doteq \Phi(u)$, onde u segue do exemplo 3.4 como $u = p(w)$ e $w = (k + 0,5 - \lambda)/\lambda^{1/2}$. Entretanto, a aproximação mais precisa para $P = P(Y \leq k)$ segue de $P \doteq \Phi(u)$ com

$$z = \Phi\left(\left\{k + \frac{2}{3} - \lambda + \frac{\epsilon}{k+1}\right\}\{1 + g[(k + 0,5)/\lambda]\}^{1/2}\,\lambda^{-1/2}\right),$$

onde $g(x)$ foi definido logo após a equação (3.41). A constante ϵ só é relevante para λ pequeno e pode ser considerada igual a 0,02 ou, se nas extremidades, igual a 0,022. Esta aproximação tem erro de ordem $\lambda^{-3/2}$ uniformemente em k, com alguma deterioração no caso trivial $k = 0$, onde $P = e^{-\lambda}$ não requer a aproximação normal.

Uma probabilidade pontual $P(Y = k)$ pode ser calculada como

$$\begin{aligned} P(Y = k) &\doteq \Phi\left(\frac{k + 0,5 - \lambda}{\sqrt{\lambda}}\right) \\ &- \Phi\left(\frac{k - 0,5 - \lambda}{\sqrt{\lambda}}\right). \end{aligned}$$

Se k é grande, através da aproximação de Stirling para $\Gamma(k + 1) = k!$, obtém-se

$$P(Y = k) = \frac{e^{k-\lambda}}{\sqrt{2\pi k}}\left(\frac{\lambda}{k}\right)^k\left\{1 + \frac{1}{12k} + O(k^{-2})\right\}.$$

Distribuição Binomial Negativa

A distribuição binomial negativa $B^-(s, p)$ é definida em ensaios independentes de Bernoulli para modelar a variável aleatória que representa o número Y de falhas verificadas antes de ocorrerem s sucessos. Então,

$$P(Y = k) = \binom{s + k - 1}{k} p^s(1 - p)^k,$$

sendo p a probabilidade de sucesso e $k = 0, 1, 2, \ldots$ Tem-se

$$P = P(Y \leq k) = \sum_{j=0}^{k} \binom{s + j - 1}{j} p^s(1 - p)^j,$$

que é idêntica a $P(X \geq s)$, sendo $X \sim B(s+k,p)$. Logo, da equação (3.39) vem

$$P(Y \leq k) \doteq \Phi\left(2[\{(k+1)p\}^{1/2} - \{s(1-p)\}^{1/2}]\right) .$$

A distribuição binomial negativa pode ser normalizada através da transformação $Z = \sqrt{s}\,\mathrm{arcsenh}(\sqrt{Y/s})$, tendo Z, aproximadamente, distribuição normal $N(0,1)$.

Distribuição Hipergeométrica

Considere uma população de N elementos classificada em S sucessos e $N-S$ fracassos. Retira-se desta população, sem reposição, uma amostra de n indivíduos. O número Y de sucessos nesta amostra tem distribuição hipergeométrica de parâmetros (S,n,N) com função de probabilidade

$$P(Y=k) = \frac{\binom{S}{k}\binom{N-S}{n-k}}{\binom{N}{n}} \tag{3.44}$$

para $k = 0,1,\ldots,\min(S,n)$. Demonstra-se que $\mu = E(Y) = np$ e $\sigma^2 = \mathrm{Var}(Y) = np(1-p)\frac{(N-n)}{(N-1)}$, onde $p = S/N$. Uma aproximação para a função de distribuição de (3.44) é dada por

$$P = P(Y \leq k) \doteq \Phi((k+0,5-\mu)/\sigma) .$$

Sejam $\tau = np(1-p)(1-\frac{n}{N})$, $w = (k+0,5-\mu)/\sigma$ e $v = (k+0,5-\mu)/\tau$. Demonstra-se que Y tem distribuição assintoticamente normal quando $N \longrightarrow \infty$ se, e somente se, $\mu \longrightarrow \infty$ e $\tau \longrightarrow \infty$. A aproximação $\Phi(v)$ para P é melhor do que $\Phi(w)$, e está correta até ordem $O(\tau^{-1})$. Uma aproximação aperfeiçoada para P, correta até $O(\tau^{-2})$, é $P \doteq \Phi(u)$, onde

$$u = v + \frac{(1-v^2)(N-2S)(N-2n)}{6N^2\tau} + \frac{v\{N^2 - 3S(N-S)\}}{48N^2\tau^2} .$$

As probabilidades pontuais (3.44) podem ser aproximadas pelas distribuições binomial e de Poisson. Usando a distribuição binomial, tem-se como primeira aproximação, quando $n < 0,1N$,

$$P(Y=k) \doteq \binom{n}{k} p^k (1-p)^{n-k} .$$

Uma melhoria nesta aproximação pode ser conseguida substituindo n e p por $n^* = np/p^*$ e $p^* = \{(n-1)+(N-n)p\}/(N-1)$. Uma aproximação assintótica cujo termo principal é a distribuição binomial é dada por

$$P(Y=k) = \binom{n}{k} p^k (1-p)^{n-k} \left[1 + \frac{\{k-(k-np)^2\}}{2Np} + O\left(\frac{1}{N^2p^2}\right)\right].$$

Se $n > Np$, uma expansão melhor é obtida permutando n e NP. Quando p é pequeno e n é grande, pode-se usar a distribuição de Poisson como aproximação tal qual

$$P(Y=k) = \frac{e^{-np}(np)^k}{k!}\left[1 + \left(\frac{1}{2Np} + \frac{1}{2n}\right)\{k-(k-np)^2\} + O\left(\frac{1}{k^2} + \frac{1}{n^2}\right)\right].$$

3.11 Exercícios

1. Calcule a função de distribuição da soma S_n de 3, 4, 5 e 6 variáveis aleatórias uniformes em $(0,1)$. Compare numericamente as probabilidades $P(S_n \geq s)$ exatas com aquelas obtidas das expansões de Edgeworth até $O(n^{-1/2})$ e $O(n^{-1})$, fazendo s igual a um, dois e três desvios padrão acima da média de S_n.

2. Seja χ'^2_{r,λ^2} uma variável aleatória qui-quadrado não-central, com r graus de liberdade e parâmetro de não-centralidade λ^2, cuja função geratriz de momentos é $M(t) = (1-2t)^{-1/(2r)} \exp\{t\lambda^2(1-2t)^{-1}\}$.

   (a) Demonstre por expansão direta que

   $$M(t) = (1-2t)^{-1/2r}\left\{1 + t\lambda^2(1-2t)^{-1} + \frac{1}{2}t^2\lambda^4(1-2t)^{-2} + O(\lambda^6)\right\};$$

   (b) Demonstre por inversão de $M(t)$, que a função densidade $f(y;r,\lambda^2)$ da variável χ'^2_{r,λ^2} pode ser expandida em termos da função densidade $f_r(y)$ de uma variável aleatória qui-quadrado central χ^2_r, com r graus de liberdade, como

   $$f(y;r,\lambda^2) = f_r(y) + \frac{\lambda^2}{2}\{f_r(y) - f_{r+2}(y)\} + \frac{\lambda^4}{8}\{f_r(y) - 2f_{r+2}(y) + f_{r+4}(y)\} + O(\lambda^6).$$

3. Sejam $Y_1, \ldots, Y_n$ variáveis aleatórias contínuas *iid* com distribuição na família exponencial (3.16) com $\lambda = \lambda_0$. Demonstre que a função densidade da soma $S_n = \sum_{i=1}^{n} Y_i$ pode ser expressa por

$$f_{S_n}(s;\lambda_0) = \frac{\exp[-(\hat{\lambda}-\lambda_0)s + n\{K(\hat{\lambda}) - K(\lambda_0)\}]}{\sqrt{2n\pi\, K''(\hat{\lambda})}}\{1 + O(n^{-1})\}\,.$$

4. Deduza a expansão de Edgeworth para a convolução S_n de n variáveis aleatórias *iid* cuja função de distribuição é $F(y) = \Phi(\sqrt{y})$, $y > 0$. Calcule numericamente as probabilidades $P(S_n \geq s)$ através das expansões de Edgeworth e da aproximação de Lugannani e Rice para $n = 5, 10$ e 20 e s igual à média de S_n e igual a 2 e 3 desvios padrão de S_n acima da média.

5. Compare numericamente as aproximações (3.25) com as correções de continuidade e (3.26) no cálculo das probabilidades $P(S_5 \geq s)$ de uma soma de 5 variáveis aleatórias *iid* com distribuição em série logarítmica, cuja função densidade é $P(Y = y;\theta) = \alpha\theta^y/y$, $\alpha = -\{\log(1-\theta)\}^{-1}$, $0 < \theta < 1$ e $y = 1, 2, \ldots$ Faça $\theta = 0,2$, $0,4$, $0,6$ e $0,8$ e $s = 5E(y) + k\sqrt{5}\mathrm{Var}(Y)^{1/2}$, onde $k = 0,1$ e 2, $E(Y) = \alpha\theta/(1-\theta)$ e $\mathrm{Var}(Y) = \alpha\theta(1-\alpha)/(1-\theta)^2$.

6. Demonstre que para a distribuição gama, cuja função densidade é $f(y) = \alpha^r y^{r-1} e^{-\alpha y}/\Gamma(y)$, tem-se

$$\lim_{r\to\infty} P\left\{\frac{(\alpha Y - r)}{\sqrt{r}} \leq y\right\} = \Phi(y)\,.$$

7. Demonstre as expansões abaixo:

(*a*) $\Gamma(n+1) = \sqrt{2\pi}(n+1)^{n+0,5}e^{-n-1}\left\{1 + \dfrac{1}{12(n+1)} + \dfrac{1}{288(n+1)^2} - \cdots\right\}$;

(*b*) $\Gamma(n+0,5) = \sqrt{2\pi}n^n e^{-n}\exp\left\{-\dfrac{1}{24n} + \dfrac{7}{2880n^3} + O(n^{-5})\right\}$.

8. Demonstre que a função de distribuição da χ^2_n pode ser expressa da expansão de

Gram-Charlier como

$$F_{\chi^2_n}(y) = \Phi(y) - \frac{\gamma_1}{6}\Phi^{(3)}(y) + \frac{\gamma_2}{24}\Phi^{(4)}(y) ,$$

onde $\gamma_1 = \sqrt{128}/(27n\sqrt{n})$, $\gamma_2 = -4/(9n) - 64/(81n^2)$ e $\Phi^{(k)}(y) = \frac{d^k\Phi(y)}{dy^k}$.

9. Calcule a expansão ponto de sela para a soma de n variáveis aleatórias binomiais $B(m_j, \mu)$ com a mesma probabilidade de sucesso μ mas com índices $m_1, \ldots, m_n$ diferentes.

Capítulo 4

Teoria Assintótica de Primeira Ordem

4.1 Fundamentos

Neste capítulo apresenta-se a *teoria assintótica de primeira ordem* definida na inferência estatística dos modelos paramétricos supondo que a informação é grande. Nesta teoria os resultados são válidos somente quando $n \longrightarrow \infty$ e decorrem de técnicas de linearização local baseadas nas expansões em série de Taylor e nos teoremas centrais do limite. Em especial, a função escore sendo uma soma de componentes independentes tem assintoticamente distribuição normal. A linearização local relaciona a distribuição da EMV com a distribuição da função escore, implicando que a EMV também tem assintoticamente distribuição normal. A teoria assintótica de primeira ordem produz uma variedade de métodos e testes estatísticos que são equivalentes somente até esta ordem mas diferem por quantidades de ordem inferior.

A teoria assintótica de primeira ordem geralmente admite que o número de observações n cresce mas a dimensão do vetor de parâmetros p se mantém constante. Ela é importante porque produz simplificações consideráveis para problemas em grandes amostras, implicando resultados simples e elegantes. Ao contrário, a teoria em pequenas amostras é extremamente complicada e as soluções exatas têm alto grau de complexidade. O ponto fundamental a favor da teoria assintótica de primeira ordem é que as soluções aproximadas mostram-se, em geral, bastante razoáveis mesmo quando n não é grande. Esta teoria

é importante por dois motivos bem distintos. O primeiro surge quando não se tem em princípio uma solução exata para o problema estatístico ou quando a solução exata é muito complicada. Então, pode ser muito mais vantajoso obter uma aproximação simples em grandes amostras para alcançar objetivos práticos ou para se ter mais informação sobre a solução exata do problema. O segundo motivo, o mais freqüente, revela o seu papel central na inferência estatística quando o problema *realmente* não tem solução exata, como, por exemplo, quando não existe uma região de confiança exata ou um teste ótimo para o parâmetro de interesse. Então, torna-se natural e inevitável obter soluções aproximadas supondo que o número de observações é grande.

Nesta seção apresentam-se alguns critérios mais comuns (erro médio quadrático e eficiência) para selecionar as estimativas dos parâmetros nos modelos estatísticos e estudam-se as propriedades assintóticas de maior interesse das EMV, tais como, consistência, unicidade, normalidade, eficiência e suficiência. Estas propriedades são válidas somente quando $n \longrightarrow \infty$ e formam a base da teoria assintótica de primeira ordem com o objetivo de se fazer inferência.

4.1.1 Erro Médio Quadrático

Considera-se aqui apenas o caso uniparamétrico ($p = 1$). O *erro médio quadrático* (EMQ) é uma das medidas preferidas para medir o desempenho de uma estimativa T de um escalar θ, sendo definido por

$$EMQ(T) = E\{(T-\theta)^2\} = \text{Var}(T) + B(\theta)^2,$$

onde $B(\theta) = E(T) - \theta$ é o viés de T. Em geral, tem-se interesse em estimativas *não-viesadas* ($B(\theta) = 0$) *de variância mínima* (NVVM) visando reduzir o EMQ. Entretanto, em muitas situações, pode-se preferir uma estimativa cujas quantidades $B(\theta)$ e $\text{Var}(T)$ são pequenas a uma outra estimativa não-viesada mas de variância apreciável. As estimativas de EMQ mínimo não são muito usadas face a dificuldades em minimizar o EMQ sem restrições adicionais. Entretanto, existe uma teoria elegante para as estimativas NVVM que tornam estas estimativas atraentes. O EMQ fornece um limite superior para a probabilidade de que o erro absoluto de estimação exceda uma determinada quantidade pois,

pela desigualdade de Chebyshev,

$$P(|T-\theta| \geq \varepsilon) \leq EMQ(T)/\varepsilon^2.$$

As EMV em geral são viesadas em pequenas amostras e na Seção 5.3 mostra-se como calcular os seus vieses de ordem n^{-1}. Entretanto, as EMV são assintoticamente não-viesadas.

4.1.2 Eficiência

É óbvio que quanto menor for a variância de uma estimativa não-viesada, maior será a chance desta estimativa estar próxima do parâmetro verdadeiro. Uma propriedade desejável é que a variância de uma estimativa não-viesada seja tão pequena quanto possível. Esta propriedade conduz a estimativas mais eficientes. Na estimação de um escalar θ, uma estimativa T é mais eficiente do que uma outra T' (no sentido de usar mais eficientemente as observações) se $EMQ(T) \leq EMQ(T')$. A *eficiência relativa* de T' em relação a T é expressa pelo quociente $e(T',T) = EMQ(T)/EMQ(T')$ e geralmente depende de θ. No caso de estimativas não-visadas, a eficiência reduz-se ao quociente das variâncias das estimativas e, então, a estimativa NVVM é a mais eficiente. Felizmente, em problemas regulares, existe um limite inferior tal que a variância de uma estimativa não pode ser menor do que este limite. Para qualquer estimativa T de um parâmetro θ cujo viés é $B(\theta)$, a sua variância satisfaz $\mathrm{Var}(T) \geq \{1+B'(\theta)\}^2/K(\theta)$, onde $B'(\theta) = dB(\theta)/d\theta$. Esta expressão é conhecida como desigualdade de Cramér-Rao. Se a estimativa é não-viesada, a variância mínima se iguala ao inverso da informação.

Se uma estimativa T tem esperança $E(T) = \tau(\theta)$, a desigualdade de Cramér-Rao passa a ser $\mathrm{Var}(T) \geq \tau'(\theta)^2/K(\theta)$. Claro que a forma anterior é um caso especial desta desigualdade. Então, a *eficiência absoluta* de uma estimativa não-viesada T de $\tau(\theta)$ é definida por $e(T) = \{\mathrm{Var}(T)K(\theta)/\tau'(\theta)^2\}^{-1}$ sendo evidentemente menor ou igual a um. Se $e(T) = 1$ a estimativa T é eficiente. Quando $\tau(\theta) = \theta$, a eficiência reduz-se a $e(T) = \{\mathrm{Var}(T)K(\theta)\}^{-1}$. *A EMV $\hat{\theta}$ de θ é assintoticamente eficiente.*

Uma condição necessária e suficiente para que uma estimativa não-viesada T de $\tau(\theta)$

seja eficiente (isto é, o limite de Cramér-Rao seja alcançado) é que a função escore seja fatorada como

$$U(\theta) = \frac{K(\theta)}{\tau'(\theta)}\{T - \tau(\theta)\} \ . \tag{4.1}$$

Caso T seja não-viesada para θ, (4.1) simplifica-se para $U(\theta) = K(\theta)(T - \theta)$. Pode-se provar ainda que existe uma estimativa T do escalar $\tau(\theta)$ de variância mínima se, e somente se, os dados têm distribuição na família exponencial uniparamétrica dada por

$$f(y;\theta) = \exp\{a(y)c(\theta) - b(\theta) + d(y)\}. \tag{4.2}$$

É fácil comprovar que as equações (4.1) e (4.2) são equivalentes.

Uma propriedade importante da EMV é que se existe uma estimativa eficiente de um escalar θ, o método de máxima verossimilhança irá produzi-la. Se T é eficiente para θ, (4.1) implica que a função escore é linear em T, ou seja, $U(\theta) = C(\theta)T + D(\theta)$. Para $\theta = \hat{\theta}$ vem $C(\hat{\theta})T + D(\hat{\theta}) = 0$. Como uma estimativa de θ eficiente é não-viesada obtém-se de $E\{U(\theta)\} = 0$: $C(\theta)\theta + D(\theta) = 0$. Avaliando esta expressão em $\hat{\theta}$, encontra-se $\hat{\theta} = T$.

Há uma correspondência biunívoca entre a existência de uma estatística suficiente para θ e a existência de uma estimativa NVVM para alguma função de θ desde que o campo de variação dos dados independa do parâmetro desconhecido. Com efeito, se S é uma estatística suficiente para θ, a equação (1.5) é válida, e derivando o seu logaritmo em relação a θ resulta na seguinte expressão para a função escore:

$$U(\theta) = \frac{\partial}{\partial\theta} \log g(s,\theta) = M(s,\theta),$$

onde M é alguma função de s e θ. Satisfeitas algumas condições de regularidade, pode-se provar que esta equação implica os dados terem distribuição na família (4.2) e, portanto, que apenas uma função desta estatística $T = T(S)$ (T é também suficiente para θ) irá satisfazer (4.1), ou seja, irá estimar alguma função $\tau(\theta)$ de θ com variância igual ao valor mínimo $\tau'(\theta)^2/K(\theta)$. No sentido inverso, quando (4.1) for satisfeita, (4.2) será verificada e, obviamente, existirá uma estatística suficiente para θ. Constata-se ainda comparando (4.1) com a equação $U(\theta) = M(s,\theta)$ que a condição de suficiência é bem menos restritiva que a condição de existência da estimativa NVVM.

Seja $\mathcal{F}$ uma certa classe de distribuições e suponha que todas as estimativas T de um parâmetro escalar θ sejam não-viesadas e cujas variâncias existem para toda distribuição desta classe. Lehmann e Scheffé (1950) mostraram que no máximo uma destas estimativas é a mais eficiente para a classe $\mathcal{F}$ em consideração. O teorema de Rao-Blackwell (Lehmann, 1983, Seção 1.6) mostra que é sempre possível a partir de uma estimativa T de θ não-viesada e de uma estimativa S de θ suficiente, construir uma outra estimativa não-viesada de θ que seja pelo menos tão eficiente quanto T. Matematicamente, a estatística $E(T|S)$ é uma estimativa não-viesada de θ e, se $\mathrm{Var}(T)$ existir, a sua variância irá satisfazer

$$\mathrm{Var}\{E(T|S)\} \leq \mathrm{Var}(T) .$$

A igualdade na expressão acima ocorrerá se $E(T|S) = T$ com probabilidade igual a um.

4.1.3 Condições de Regularidade

As condições seguintes de regularidade são usadas na teoria assintótica para justificar e delimitar os erros das expansões em série de Taylor. Algumas dessas condições ou a totalidade delas são necessárias para provar as propriedades assintóticas das EMV de consistência, unicidade, normalidade, eficiência e suficiência, apresentadas nas Seções 4.1.4 - 4.1.7 e 4.2.

Suponha que os dados $y_i's$ são realizações *iid* de uma variável aleatória Y caracterizada por distribuições P_θ pertencentes a uma certa classe $\mathcal{P}$, que dependem de um vetor θ de dimensão p, $\theta \in \Theta$. Sejam $f(y;\theta)$ e $L(\theta) = \Pi f(y_i;\theta)$ as funções de probabilidade ou densidade comum dos dados e de verossimilhança para θ, respectivamente.

As seguintes suposições serão necessárias no decorrer deste capítulo:

(i) as distribuições P_θ são identificáveis, isto é, $\theta \neq \theta' \in \Theta$ implica $P_\theta \neq P_{\theta'}$;

(ii) as distribuições P_θ têm o mesmo suporte para todo $\theta \in \Theta$, ou seja, o conjunto $A = \{y; f(y;\theta) > \theta\}$ independe de θ.

A condição (i) assegura que as distribuições de probabilidade dos dados definidas por dois valores distintos de θ são diferentes e a condição (ii) garante que seus campos de variação são idênticos e independem de θ. As suposições (iii) - (v) abaixo garantem a

regularidade de $f(y;\theta)$ como função de θ e a existência de um conjunto aberto Θ_1 no espaço paramétrico Θ tal que o parâmetro verdadeiro θ_0 pertença a Θ_1:

(iii) existe um conjunto aberto Θ_1 em Θ contendo θ_0 tal que a função densidade $f(y;\theta)$, para quase todo y, admite todas as derivadas até terceira ordem em relação a θ, para todo $\theta \in \Theta_1$;

(iv) $E_\theta\{U(\theta)\} = 0$ e a matriz de informação $0 < K(\theta) < \infty$ para todo $\theta \in \Theta_1$;

(v) existem funções $M_{ijk}(y)$ independentes de θ tais que, para $i, j, k = 1, \ldots, p$,

$$\left|\frac{\partial^3 \log f(y;\theta)}{\partial\theta_i \partial\theta_j \partial\theta_k}\right| < M_{ijk}(y)$$

para todo $\theta \in \Theta_1$, onde $E_{\theta_0}\{M_{ijk}(Y)\} < \infty$.

A condição (iii) representa a existência de Θ_1 e de derivadas de $f(y;\theta)$ até terceira ordem em Θ_1, a (iv) que a matriz de informação é finita e positiva definida numa vizinhança aberta de θ_0 e a (v) que as terceiras derivadas da log-verossimilhança são limitadas por uma função integrável de Y cuja esperança é finita.

4.1.4 Consistência

Usualmente, uma estimativa é função (explícita ou implícita) do tamanho da amostra n e, pelo menos intuitivamente, espera-se que a precisão desta estimativa aumente quando $n \longrightarrow \infty$. Neste sentido, uma estimativa T_n é chamada de *consistente* para um parâmetro θ se $EMQ(T_n) \longrightarrow 0$ quando $n \longrightarrow \infty$. A grande maioria dos métodos de estimação, como o método de máxima verossimilhança, produz estimativas consistentes segundo certas condições de regularidade. Geralmente, duas definições de consistência são usadas amplamente na teoria assintótica. Sejam estimativas T_n baseadas em variáveis aleatórias *iid*, $Y_1, \ldots, Y_n$. Diz-se que T_n é: (a) *fracamente consistente* para θ se $T_n = \theta + o_p(1)$; (b) *fortemente consistente* para θ se $T_n = \theta + o(1)$ com probabilidade um. A consistência fraca (forte) ocorre quando T_n satisfaz à lei fraca (forte) dos grandes números. Então, T_n é fracamente ou fortemente consistente para θ se $\lim_{n\to\infty} P_\theta(|T_n - \theta| \geq \varepsilon) = 0$, $\forall \varepsilon > 0$ ou $P_\theta(\lim_{n\to\infty} T_n = \theta) = 1$, respectivamente. Uma propriedade importante da EMV é a consistência (forte) supondo válidas algumas condições de regularidade da Seção 4.1.3.

Para n fixo define-se a EMV $\hat{\theta}$ em Θ de modo que

$$\ell(\hat{\theta}) \geq \ell(\theta) \tag{4.3}$$

para todo $\theta \in \Theta$. Por causa da igualdade em (4.3), a seqüência de valores de $\hat{\theta}$ quando $n \longrightarrow \infty$ poderá não ser univocamente determinada. Mostra-se aqui que se as condições de regularidade (i) – (ii) da Seção 4.1.3 são válidas e Θ é finito, então a EMV $\hat{\theta}$ é (fortemente) consistente para o parâmetro verdadeiro $\theta_0 (\hat{\theta} \xrightarrow{q.c.} \theta_0)$, ou seja, $P_{\theta_0}\left(\lim_{n\to\infty} \hat{\theta} = \theta_0\right) = 1$. Uma versão simplificada da demonstração usa a desigualdade de Jensen – $E\{\psi(Z)\} \leq \psi(E(Z))$ – válida quando $\psi(Z)$ é uma função côncava definida em $\mathbb{R}$ e Z é uma variável aleatória integrável. Como a função logaritmo é estritamente côncava, pode-se aplicá-la à variável aleatória $L(\theta)/L(\theta_0)$ para obter

$$E_0\left[\log\left\{\frac{L(\theta)}{L(\theta_0)}\right\}\right] < \log\left[E_0\left\{\frac{L(\theta)}{L(\theta_0)}\right\}\right] ,$$

para todo $\theta \neq \theta_0$, onde E_0 significa o operador esperança matemática segundo o parâmetro θ_0. Mas $E_0\{L(\theta)/L(\theta_0)\} = 1$ e, portanto, $E_0\{\ell(\theta)\} < E_0\{\ell(\theta_0)\}$ para todo $\theta \neq \theta_0$. A essência da demonstração da consistência de $\hat{\theta}$ é que (4.3) e $E_0\{\ell(\theta)\} < E_0\{\ell(\theta_0)\}$ são incompatíveis a menos que $\hat{\theta}$ convirja para θ_0. Pela lei (forte) dos grandes números $n^{-1}\ell(\theta) = n^{-1}\Sigma \log f(y_i;\theta)$ converge para $n^{-1}E_0\{\ell(\theta)\}$ quando $n \longrightarrow \infty$. Logo, por causa de $E_0\{\ell(\theta)\} < E_0\{\ell(\theta_0)\}$ vem

$$\lim_{n\to\infty} P_{\theta_0}(\ell(\theta) < \ell(\theta_0)) = 1, \quad \theta \neq \theta_0 . \tag{4.4}$$

O limite em (4.4) especifica que, para n grande, a log-verossimilhança em θ_0 excede o seu valor em qualquer outro ponto $\theta \neq \theta_0$, com probabilidade próxima de um. Os resultados (4.4) e (4.3) com $\theta = \theta_0$ só não serão incompatíveis para n grande se $P_{\theta_0}\left(\lim_{n\to\infty} L(\hat{\theta}) = L(\theta_0)\right) = 1$ for satisfeita. As condições (i) – (ii) e a finitude de Θ permitem concluir que $P_{\theta_0}\left(\lim_{n\to\infty} \hat{\theta} = \theta_0\right) = 1$, ou seja, $\hat{\theta}$ é (fortemente) consistente para θ_0.

Se Θ for infinito ou mesmo infinito enumerável não se pode deduzir a consistência

(forte) de $\hat{\theta}$ diretamente de (4.4) sem as suposições (iii) - (v) da seção anterior. Assim, prova-se agora a *consistência* (forte) da EMV $\hat{\theta}$, supondo que as condições de regularidade (i) - (v) são satisfeitas, a partir do resultado (4.4) na situação geral de Θ infinito. Como a log-verossimilhança é diferenciável por (iii), obtém-se por expansão de $\ell(\hat{\theta})$ em série de Taylor até segunda ordem

$$\ell(\hat{\theta}) = \ell(\theta_0) + U(\theta_0)(\hat{\theta} - \theta_0) - \frac{1}{2}(\hat{\theta} - \theta_0)^T J(\theta^*)(\hat{\theta} - \theta_0) \tag{4.5}$$

onde $J(\theta)$ é a informação observada (Seção 1.3) para θ e θ^* é um vetor situado entre $\hat{\theta}$ e θ_0. Como $U(\theta_0)$ e $J(\theta^*)$ são somas de variáveis aleatórias *iid* elas convergem pela lei (forte) dos grandes números para os seus respectivos valores esperados. Logo, por (iv), $U(\theta) \xrightarrow{q.c.} 0$ e $J(\theta^*) \xrightarrow{q.c.} K(\theta^*) > 0$, e de (4.2) com $\theta = \theta_0$ mais (4.4) conclui-se que $\ell(\hat{\theta}) \xrightarrow{q.c.} \ell(\theta_0)$. Deste modo, a forma quadrática em (4.5) deve aproximar-se de zero quando n cresce e, forçosamente, $\hat{\theta} \xrightarrow{q.c.} \theta_0$. Então, demonstrou-se a *consistência* (forte) de qualquer seqüência de estimativas $\hat{\theta}$ obtidas segundo (4.3).

Segundo as condições (i) - (v) pode-se também demonstrar que, com probabilidade tendendo a um quando $n \longrightarrow \infty$, existe pelo menos uma seqüência de soluções $\tilde{\theta}$ da equação de máxima verossimilhança $U(\theta) = 0$ tal que $\tilde{\theta} \xrightarrow{q.c.} \theta_0$, ou seja, $\tilde{\theta}$ *é fortemente consistente* para θ_0. A prova formal, entretanto, é bastante complicada e será omitida aqui. Se as observações forem independentes mas não identicamente distribuídas, muitos dos argumentos usados anteriormente continuarão valendo aplicando-se a lei fraca dos grandes números.

4.1.5 Unicidade Assintótica

Segundo as condições gerais (i) - (v) pode-se demonstrar a *unicidade assintótica* de $\hat{\theta}$, isto é, para n grande existe uma única EMV de θ_0. Em outras palavras, para grandes amostras a log-verossimilhança se torna estritamente côncava. Antes de demonstrar a existência de uma única EMV para θ_0 quando $n \longrightarrow \infty$ mostra-se que para n grande $\hat{\theta}$ é uma solução da equação de MV, $U(\theta) = 0$ e, com probabilidade um, corresponde a um máximo local em qualquer conjunto aberto centrado em θ_0. Expandindo $U(\theta)$ até

primeira ordem e fazendo $\theta = \hat{\theta}$ vem, com a mesma notação de (4.5),

$$U(\hat{\theta}) = U(\theta_0) - J(\theta^*)(\hat{\theta} - \theta_0)\,. \tag{4.6}$$

Os dois termos no lado direito de (4.6) tendem a zero quando $n \longrightarrow \infty$; o primeiro pela lei forte dos grandes números e o segundo pela consistência da EMV. Logo, para n grande, $\hat{\theta}$ é uma solução de $U(\theta) = 0$. Como as observações são *iid* pode-se considerar $K(\theta) = nk(\theta)$, para todo $\theta \in \Theta$, onde $k(\theta)$ (> 0 por (iv)) é a matriz de informação para θ relativa a uma única observação. Pela consistência forte de $\hat{\theta}$ vem $n^{-1}J(\hat{\theta}) \xrightarrow{q.c.} n^{-1}J(\theta_0)$ e, pela lei forte dos grandes números, $n^{-1}J(\theta_0)$ converge com probabilidade um para $k(\theta_0) > 0$. A conjunção dos dois resultados implica que qualquer EMV $\hat{\theta}$ deve verificar

$$\lim_{n\to\infty} P_{\theta_0}(J(\hat{\theta}) > 0) = 1, \tag{4.7}$$

de onde se conclui que $\hat{\theta}$ corresponde, com probabilidade um, a um máximo local de $U(\theta) = 0$. Prova-se agora facilmente a *unicidade assintótica* de $\hat{\theta}$. Para n grande, se (4.3) produzisse duas EMV $\hat{\theta}'$ e $\hat{\theta}''$, elas seriam consistentes e verificariam $U(\theta) = 0$ e (4.7), ou seja, seriam máximos locais assintoticamente. Então, existiria entre $\hat{\theta}'$ e $\hat{\theta}''$ um ponto de mínimo $\overline{\theta}$ consistente para $\theta_0(\overline{\theta} \xrightarrow{q.c.} \theta_0)$ satisfazendo $J(\overline{\theta}) < 0$. Mas isto violaria (iv), pois para n grande, $J(\theta)$ deve ser positiva definida para $\theta \in \Theta_1$. Como a ocorrência de dois máximos locais consistentes implica uma contradição fica provada a unicidade da EMV $\hat{\theta}$ em grandes amostras.

Em geral, no caso multiparamétrico $p \geq 2$, mesmo que $U(\theta) = 0$ tenha solução única não implica que ela seja a EMV de θ que pode até mesmo nem existir. Contudo, no caso uniparamétrico ($p = 1$), se a solução da equação de MV for única, a probabilidade de que esta solução seja a EMV tenderá para um quando $n \longrightarrow \infty$. Haverá unicidade das equações de MV quando $f(y;\theta)$ for uma distribuição não-degenerada pertencente à família exponencial com p parâmetros (Seção 1.5), pois $\ell(\theta)$ será estritamente côncava.

4.1.6 Normalidade Assintótica

Considere n observações *iid*, supondo válidas as condições de regularidade (i) - (v) da Seção 4.1.3. Se $\tilde{\theta}$ é uma solução consistente da equação de máxima verossimilhança $U(\theta) = 0$, então

$$\sqrt{n}(\tilde{\theta} - \theta_0) \xrightarrow{\mathcal{D}} N_p(0, k(\theta_0)^{-1}), \tag{4.8}$$

ou seja, em grandes amostras, a distribuição de $\tilde{\theta}$ é aproximadamente normal p-dimensional com média θ_0 e matriz de covariância $K(\theta_0)^{-1} = n^{-1}k(\theta_0)^{-1}$. Cramér (1946, Seção 33.3) e Lehmann (1935, Seção 6.4) apresentam demonstrações rigorosas da convergência (4.8) para $p = 1$ e $p \geq 1$, respectivamente. Mostra-se inicialmente a demonstração de (4.8) no caso uniparamétrico. As condições gerais de regularidade garantem a expansão de $U(\tilde{\theta}) = 0$ em torno do parâmetro verdadeiro θ_0 até segunda ordem:

$$U(\theta_0) + U'(\theta_0)(\tilde{\theta} - \theta_0) + \frac{1}{2}U''(\theta^*)(\tilde{\theta} - \theta_0)^2 = 0,$$

onde $|\theta^* - \theta_0| < |\tilde{\theta} - \theta_0|$ e, portanto, θ^* é necessariamente consistente para θ_0. Os dois primeiros termos no lado esquerdo desta equação são $O_p(n^{1/2})$ e o terceiro é $O_p(1)$, pois $U'(\theta_0) = O_p(n)$, $U''(\theta^*) = O_p(n)$ e $\tilde{\theta} - \theta_0 = O_p(n^{-1/2})$. Como $U(\theta_0)$ e $U'(\theta_0)$ são somas de variáveis aleatórias *iid*, a expansão anterior implica

$$\sqrt{n}(\hat{\theta} - \theta_0)\left\{\frac{-\sum_{i=1}^{n} U_i'(\theta_0)}{nk(\theta_0)} + O_p(n^{-1/2})\right\} = \frac{\sum_{i=1}^{n} U_i(\theta_0)}{\sqrt{n}k(\theta_0)}.$$

Pela lei fraca dos grandes números $-\sum_{i=1}^{n} U_i'(\theta_0)/\{nk(\theta_0)\} = 1 + o_p(1)$. Logo,

$$\sqrt{n}(\hat{\theta} - \theta_0)\{1 + o_p(1)\} = \frac{\sum_{i=1}^{n} U_i(\theta_0)}{\sqrt{n}k(\theta_0)}. \tag{4.9}$$

Observe-se que (4.9) é o caso uniparamétrico da aproximação (1.15), a última equação sem o erro estocástico. Aplicando o teorema central do limite à soma estocástica do lado direito de (4.9) e por (iv) prova-se a convergência (4.8).

A demonstração da *normalidade assintótica* de $\tilde{\theta}$ no caso multiparamétrico é feita de forma análoga ao caso $p = 1$. Quando θ for um vetor de dimensão p, a igualdade (4.9) é generalizada para

$$\sqrt{n}(\hat{\theta} - \theta_0)\{1 + o_p(1)\} = \frac{1}{\sqrt{n}} k(\theta_0)^{-1} U(\theta_0), \tag{4.10}$$

onde $k(\theta) = n^{-1}K(\theta)$ é a matriz de informação para uma única observação. De (4.10) e (iv) é fácil checar que $\hat{\theta}$ tem média assintótica zero e estrutura de covariância assintótica dada por $\mathrm{Cov}(\tilde{\theta}) = K(\theta_0)^{-1}$. Então, a normalidade p-dimensional assintótica de $\tilde{\theta}$ decorre do teorema central do limite multivariado aplicado ao termo do lado direito de (4.10).

O fato de se aproximar a distribuição da EMV $\tilde{\theta}$ por $N_p(\theta_0,\ n^{-1}k(\theta_0)^{-1})$ é um dos resultados mais relevantes da teoria assintótica de primeira ordem com objetivos de inferência.

4.1.7 Eficiência Assintótica

No caso $p = 1$, observe-se que $k(\theta_0)^{-1}$ é a variância da distribuição assintótica de $\sqrt{n}(\tilde{\theta}-\theta_0)$ que, em geral, não coincide com o limite de Cramér-Rao (Seção 4.1.2) para a sua variância exata. Este fato é melhor compreendido observando que para qualquer estimativa T de θ assintoticamente normal, i.e.,

$$\sqrt{n}(T - \theta) \xrightarrow{\mathcal{D}} N(0, v(\theta)),\ v(\theta) > 0, \tag{4.11}$$

tem-se: $\lim_{n\to\infty}\{n\,\mathrm{Var}(T)\} \geq v(\theta) \geq k(\theta)^{-1}$. O resultado (4.11) implica que a estimativa T é consistente para θ, mas ela pode ter viés não-nulo (para n finito). Contrariamente, o limite de Cramér-Rao $k(\theta)^{-1}$ é relativo à variância exata de $\sqrt{n}(T - \theta)$ exigindo-se que ela seja necessariamente não-viesada. Uma estimativa T de θ é *assintoticamente eficiente* se satisfaz (4.11) com $v(\theta) = k(\theta)$. Desta definição e de (4.8) conclui-se que qualquer solução consistente $\tilde{\theta}$ de $U(\theta) = 0$ é assintoticamente eficiente.

Não há dificuldade em generalizar o limite de Cramér-Rao e (4.11) para as componentes de um vetor de parâmetros $\theta \in \mathbb{R}^p$. Assim, se $k(\theta)^{r,r}$ representa o r-ésimo elemento da diagonal da matriz $k(\theta)^{-1}$, $n^{-1}k(\theta)^{r,r}$ é um limite inferior para a variância assintótica de qualquer estimativa de θ_r assintoticamente normal (mesmo viesada para n

finito). A desigualdade de Cramér-Rao estabelece que qualquer estimativa não-viesada de θ_r tem variância (exata) superior a $n^{-1}k(\theta)^{r,r}$. Como por (4.8) $\lim_{n\to\infty}\{n\,\mathrm{Var}(\tilde{\theta}_r)\} = k(\theta)^{r,r}$, deduz-se que qualquer componente de $\tilde{\theta}$ é assintoticamente eficiente para o parâmetro correspondente.

Os resultados de normalidade e eficiência assintóticas apresentados aqui poderão ser generalizados para situações menos restritivas em que as observações são independentes mas não identicamente distribuídas, desde que: a) a lei fraca dos grandes números se aplique à informação observada média $n^{-1}J(\theta)$ com esta convergindo em probabilidade para $n^{-1}K(\theta)$ (a matriz de informação média); b) o teorema central do limite se aplique à função escore total $U(\theta)$ sendo a convergência para uma distribuição assintótica não-singular. Existem inúmeros outros aperfeiçoamentos com suposições mais fracas para garantir consistência, unicidade, normalidade e eficiência da EMV em situações gerais e específicas que não serão citados aqui.

4.2 Suficiência Assintótica

A fatoração de Neyman-Fisher (1.5) representa a melhor forma de se verificar a suficiência de uma estatística $S = S(Y)$. Para demonstrar a *suficiência assintótica* de uma solução $\tilde{\theta}$ da equação de máxima verossimilhança $U(\theta) = 0$ deve-se supor que as condições (i) - (v) da Seção 4.1.3 são verdadeiras. Neste caso pode-se expandir $\ell(\theta)$ analogamente à equação (4.5) como

$$\ell(\theta) = \ell(\tilde{\theta}) - \frac{1}{2}(\theta - \tilde{\theta})^T J(\tilde{\theta})(\theta - \tilde{\theta}) + o_p(1)$$

com $\theta - \tilde{\theta} = O_p(n^{-1/2})$. Portanto, a forma da verossimilhança

$$L(\theta) = L(\hat{\theta})\exp\left\{-\frac{1}{2}(\theta - \tilde{\theta})^T J(\tilde{\theta})(\theta - \tilde{\theta}) + o_p(1)\right\}$$

implica que $\tilde{\theta}$ é assintoticamente suficiente para θ, quando existir uma estatística suficiente.

Em pequenas amostras, a solução $\tilde{\theta}$ da equação de máxima verossimilhança pode não ser suficiente para θ mas sempre será função de uma estatística suficiente para θ, quando existir uma estatística suficiente.

Os resultados assintóticos deduzidos nas Seções 4.1.4 – 4.1.8 enaltecem que a teoria assintótica de primeira ordem é simples e elegante para as estimativas de máxima verossimilhança.

4.3 Inferência sem Parâmetros de Incômodo

Seja Y uma variável aleatória com função de probabilidade ou função densidade $f(y;\theta)$, $\theta \in \Theta$, dependendo de um vetor θ de dimensão p. Seja $y = (y_1, \ldots, y_n)^T$ um vetor de realizações de Y. Quando não há parâmetros de perturbação o interesse é testar a hipótese nula simples $H : \theta = \theta^{(0)}$ versus $A : \theta \neq \theta^{(0)}$, onde $\theta^{(0)}$ é um vetor especificado para θ. Há várias maneiras de testar H que são equivalentes até primeira ordem, i.e., baseiam-se em estatísticas que diferem tipicamente por quantidades de ordem $O_p(n^{-1/2})$.

Sejam $\ell(\theta), U(\theta), J(\theta)$ e $K(\theta)$ a log-verossimilhança, a função escore e as informações observada e esperada relativas ao vetor θ, respectivamente. As três estatísticas comumentes usadas para testar H versus A são a razão de verossimilhança de Neyman e Pearson $w = -2\log \ell_R$ expressa por

$$w = 2\{\ell(\hat{\theta}) - \ell(\theta^{(0)})\}, \tag{4.12}$$

a estatística escore de Rao

$$S_R = U(\theta^{(0)})^T K(\theta^{(0)})^{-1} U(\theta^{(0)}), \tag{4.13}$$

e a estatística de Wald

$$W = (\hat{\theta} - \theta^{(0)})^T K(\hat{\theta})(\hat{\theta} - \theta^{(0)}). \tag{4.14}$$

As três estatísticas acima representam as técnicas mais importantes para avaliação e teste de modelos estatísticos. A forma (4.12) foi proposta por Wilks em 1938. Depois, Wald propôs (4.14) em 1943 e Rao desenvolveu (4.13) em 1947.

As formas quadráticas (4.13) e (4.14) são deduzidas das distribuições assintóticas $N_p(0, K(\theta^{(0)}))$ e $N_p(\theta^{(0)}, K(\theta^{(0)})^{-1})$ de $U(\theta^{(0)})$ e $\hat{\theta}$, respectivamente. As estatísticas (4.12) e (4.13) independem da parametrização adotada para $f(y;\theta)$ enquanto a estatística de Wald depende da parametrização do modelo. Apenas a estatística S_R não requer o cálculo da EMV $\hat{\theta}$ embora envolva a inversa da matriz de informação.

Se as condições de regularidade (i) – (v) da Seção 4.2 são satisfeitas, as três estatísticas acima são equivalentes até primeira ordem, isto é, elas têm segundo a hipótese nula H a mesma distribuição assintótica χ^2 com p graus de liberdade. Assim, a hipótese H será rejeitada para valores grandes de w, S_R e W comparados com o valor crítico $\chi^2_p(\alpha)$ obtido da distribuição χ^2_p para um nível de significância nominal α fixado. As regiões de $100(1-\alpha)\%$ de confiança para θ são formadas, aproximadamente, por

$$R(\theta) = \{\theta; T(\theta) \leq \chi^2_p(\alpha)\},$$

onde $T(\theta)$ pode ser qualquer uma das estatísticas (4.12) – (4.14).

Como $U(\theta^{(0)})$ e $K(\theta^{(0)})$ se referem a um vetor de dados de dimensão n, pode-se adotar, sujeito a condições de estabilidade, nos cálculos assintóticos quando $n \longrightarrow \infty$, a seguinte convenção:

$$\begin{aligned} U(\theta^{(0)}) &= \sqrt{n}\,\overline{U}(\theta^{(0)}) = O_p(n^{1/2}), \\ K(\theta^{(0)}) &= n\,\overline{K}(\theta^{(0)}), \\ \hat{\theta} - \theta^{(0)} &= O_p(n^{-1/2}), \end{aligned}$$

onde $\overline{K}(\theta^{(0)})$ é a informação média por observação e $\overline{U}(\theta^{(0)})$ é a função escore normalizada. Tem-se $\overline{K}(\theta^{(0)}) = O(1)$ e $\overline{U}(\theta^{(0)}) = O_p(1)$. A vantagem da notação acima é expressar todas as quantidades em termos de outras que são de ordem $O(1)$ ou de variáveis aleatórias que são $O_p(1)$. Se as observações são *iid*, então $\overline{K}(\theta^{(0)})$ é a informação relativa a uma única observação.

Se $K(\theta)$ é contínua em $\theta = \theta^{(0)}$ obtém-se, quando $n \longrightarrow \infty$,

$$\begin{aligned} n^{-1}J(\theta^{(0)}) &\xrightarrow{p} \overline{K}(\theta^{(0)}), \\ n^{-1}J(\hat{\theta}) &\xrightarrow{p} \overline{K}(\theta^{(0)}). \end{aligned} \tag{4.15}$$

Assim, nas estatísticas (4.13) e (4.14) as matrizes $K(\theta^{(0)})$ e $K(\hat{\theta})$ podem ser substituídas pelas matrizes $J(\theta^{(0)})$ ou $J(\hat{\theta})$, pois as várias estatísticas modificadas serão equivalentes até primeira ordem, ou seja, terão a mesma distribuição limite χ^2_p. As estatísticas (4.12) – (4.14) irão diferir quando $\theta = \theta^{(0)}$ por quantidades de ordem $O_p(n^{-1/2})$.

A distribuição assintótica das estatísticas (4.12) – (4.14) é uma conseqüência da distri-

buição normal p-dimensional assintótica da função escore $U(\theta)$ com média zero e estrutura de covariância $K(\theta)$. Para observações independentes este resultado decorre da aplicação de um teorema central do limite à soma estocástica $U(\theta)$. Supõe-se aqui problemas regulares com a validade dos seguintes resultados

$$\begin{aligned} \sqrt{n}\,U(\theta) &\xrightarrow{\mathcal{D}} N_p(0, \overline{K}(\theta)), \\ \sqrt{n}(\hat{\theta}-\theta) &\xrightarrow{\mathcal{D}} N_p(0, \overline{K}(\theta)^{-1}) \,. \end{aligned} \tag{4.16}$$

A distribuição assintótica das estatísticas escore (4.13) e Wald (4.14) segue de imediato das convergências em (4.16). Para demonstrar a distribuição assintótica da razão de verossimilhança, expande-se $\ell(\theta)$ em série de Taylor em torno da solução $\hat{\theta}$ de $U(\hat{\theta}) = 0$. Assim,

$$\ell(\theta) = \ell(\hat{\theta}) - \frac{1}{2}(\theta-\hat{\theta})^T J(\hat{\theta})(\theta-\hat{\theta}) + o_p(1)$$

ou

$$w = (\hat{\theta}-\theta)^T J(\hat{\theta})(\hat{\theta}-\theta) + o_p(1). \tag{4.17}$$

Usando $\hat{\theta} - \theta = K(\theta)^{-1}U(\theta) + o_p(n^{-1/2})$ e a segunda convergência em (4.9) encontra-se

$$w = U(\theta)^T K(\theta)^{-1} U(\theta) + o_p(1) \,. \tag{4.18}$$

Usando a primeira relação de convergência em (4.16) obtém-se de (4.18) que $w \xrightarrow{\mathcal{D}} \chi^2_p$ supondo $H : \theta = \theta^{(0)}$ verdadeira. De (4.17) e (4.18) verifica-se que W e S_R são assintoticamente equivalentes a w. A mesma equivalência assintótica ocorre, pela combinação dos resultados (4.15) – (4.16), com formas variantes das estatísticas (4.13) e (4.14) deduzidas substituindo $K(\theta^{(0)})$ e $K(\hat{\theta})$ por $J(\theta^{(0)})$ ou $J(\hat{\theta})$. Assim, as estatísticas

$$\begin{gathered} U(\theta^{(0)})^T K(\hat{\theta})^{-1} U(\theta^{(0)}),\ U(\theta^{(0)})^T J(\theta^{(0)})^{-1} U(\theta^{(0)}), \\ U(\theta^{(0)})^T J(\hat{\theta})^{-1} U(\theta^{(0)}),\ (\hat{\theta}-\theta^{(0)})^T J(\theta^{(0)})(\hat{\theta}-\theta^{(0)}), \\ (\hat{\theta}-\theta^{(0)})^T J(\hat{\theta})(\hat{\theta}-\theta^{(0)}) \text{ e } (\hat{\theta}-\theta^{(0)})^T K(\theta^{(0)})(\hat{\theta}-\theta^{(0)}) \end{gathered}$$

são assintoticamente equivalentes à distribuição χ^2_p.

Exemplo 4.1 *Considere a distribuição multinomial* $y \sim M(n, \pi)$ *onde* $y = (y_1, \ldots, y_p)^T$ *(com* $y_i > 0$*) representa* p *freqüências observadas com probabilidades associadas* $\pi =$

$(\pi_1,\ldots,\pi_p)^T$. *Para testar* $H:\pi=\pi^{(0)}$ *versus* $A:\pi\neq\pi^{(0)}$, *as três estatísticas reduzem-se a:* $w=2\sum_{i=1}^{p}y_i\log\{y_i/(n\pi_i^{(0)})\}$, $S_R=\sum_{i=1}^{p}\frac{(y_i-n\pi_i^{(0)})^2}{n\pi_i^{(0)}}$ *e* $W=\sum_{i=1}^{p}(y_i-n\pi_i^{(0)})^2/y_i$. *A distribuição limite destas estatísticas segundo* H *é* χ^2_{p-1}. *A estatística* S_R *é a famosa estatística* χ^2 *de Pearson. Sobre ela R.L. Plackett foi enfático: "Pearson's chi-squared test is one of the great monuments of twentieth-century statistics."*

No caso de θ ser um escalar, as formas (4.13) e (4.14) reduzem-se a $S_R = U(\theta^{(0)})^2/K(\theta^{(0)})$ e $W=(\hat\theta-\theta^{(0)})^2K(\hat\theta)$. Buse (1982) apresenta uma excelente interpretação geométrica das formas de w, S_R e W no caso de θ escalar. Na prática, testes envolvendo um grau de liberdade podem ter mais significado comparando-se as estatísticas $\sqrt{w}, \sqrt{S_R}$ ou $\sqrt{W}$, com um sinal adequado, com os valores críticos da sua distribuição normal $N(0,1)$ assintótica. As estatísticas sinalizadas abaixo

$$\begin{aligned} r_w &= \text{sinal}(\hat\theta-\theta^{(0)})w^{1/2},\\ r_{S_R} &= U(\theta^{(0)})/\sqrt{K(\theta^{\theta})},\\ r_W &= (\hat\theta-\theta^{(0)})\sqrt{K(\hat\theta)} \end{aligned} \tag{4.19}$$

têm, assintoticamente, segundo $H:\theta=\theta^{(0)}$, distribuição normal $N(0,1)$. Aqui, também, $J(\hat\theta)$ e $J(\theta^{(0)})$ podem substituir $K(\hat\theta)$ e $K(\theta^{(0)})$ e a distribuição normal $N(0,1)$ assintótica continua valendo. Na realidade, todas estas estatísticas sinalizadas satisfazem: (i) $r=\tilde Z+O_p(n^{-1/2})$, onde Z é uma variável aleatória que tem assintoticamente distribuição normal $N(0,1)$; (ii) $P(r\leq x)=\Phi(x)+O(n^{-1/2})$. Assim, elas diferem por quantidades estocásticas de ordem $n^{-1/2}$ em probabilidade.

Exemplo 4.2 *Considere uma amostra aleatória de tamanho n da distribuição de Poisson* $P(\mu)$, *onde se deseja testar* $H:\mu=\mu^{(0)}$ *versus* $A:\mu\neq\mu^{(0)}$. *De (4.12) - (4.14) é fácil obter* $w=2n(\mu^{(0)}-\overline{y})+2n\overline{y}\log(\overline{y}/\mu^{(0)})$, $S_R=n(\overline{y}-\mu^{(0)})^2/\mu^{(0)}$ *e* $W=n(\overline{y}-\mu^{(0)})^2/\overline{y}$, *sendo* $\overline{y}$ *a média amostral. Claramente, o teste de* H *via* S_R *é equivalente ao teste baseado na aproximação normal* $N(n\mu^{(0)},n\mu^{(0)})$ *da distribuição de* $n\overline{y}$. *Qualquer uma destas estatísticas tem assintoticamente distribuição* χ^2_1.

Exemplo 4.3 *Seja uma amostra de observações* $y_1,\ldots,y_n$ *iid da distribuição exponencial*

com função densidade $f(y;\rho) = \rho e^{-\rho y}$. *A EMV de* ρ *é* $\hat{\rho} = 1/\overline{y}$. *Para testar* $H : \rho = \rho^{(0)}$ *as estatísticas em (4.19) são*

$$r_w = \text{sinal}(1 - \rho^{(0)}\overline{y})[2n\{\rho^{(0)}\overline{y} - \log(\rho^{(0)}\overline{y}) - 1\}]^{1/2},$$
$$r_{S_R} = r_W = \sqrt{n}(1 - \rho^{(0)}\overline{y}) \ .$$

Uma outra estatística equivalente é a Wald modificada r'_W *com a informação sendo avaliada na hipótese nula ao invés de ser calculada na EMV. Tem-se* $r'_W = \sqrt{n}\{(\rho_0\overline{y})^{-1} - 1\}$. *Pode ser demonstrado por primeiros princípios que*

$$r_w = -\tilde{Z} + \tfrac{1}{3\sqrt{n}}\tilde{Z}^2 + O_p(n^{-1}),$$
$$r_{S_R} = r_W = -\tilde{Z},\ r'_W = -\tilde{Z} + \tfrac{1}{\sqrt{n}}\tilde{Z}^2 + O_p(n^{-1}),$$

o que ilustra a equivalência até primeira ordem destas estatísticas, isto é, todas elas convergem em distribuição para a normal $N(0,1)$ *quando* $n \to \infty$.

As estatísticas em (4.19) são *quantidades pivotais assintóticas* para o parâmetro θ pois convergem para uma distribuição conhecida que não envolve este parâmetro quando $n \longrightarrow \infty$. Assim, os limites de $100(1-\alpha)\%$ de confiança para o escalar θ podem, alternativamente, ser obtidos como $R(\theta) = \{\theta; |r(\theta)| \leq z_\alpha\}$, onde z_α é tal que $\Phi(z_\alpha) = 1 - \alpha/2$. A estatística $r_W = (\hat{\theta} - \theta)K(\hat{\theta})^{1/2}$ tem a vantagem de englobar conjuntamente uma estimativa de θ e sua precisão $K(\hat{\theta})^{1/2}$, enquanto que a estatística de Wald alternativa $r_{W_1} = (\hat{\theta} - \theta)J(\hat{\theta})^{1/2}$, equivalente assintoticamente a r_W, contém uma variável aleatória $J(\hat{\theta})$ que não envolve θ mas pode não representar uma variância em pequenas amostras. Ambas estatísticas são lineares em θ. Quando o viés $B(\theta)$ de ordem n^{-1} de $\hat{\theta}$ (vide Seção 5.3) é apreciável, deve-se aplicar a r_W e r_{W_1} uma correção de viés (supondo $K(\hat{\theta})$ e $J(\hat{\theta})$ praticamente constantes) substituindo $\hat{\theta}-\theta$ por $\hat{\theta}-B(\hat{\theta})-\theta$. Alternativamente, determina-se intervalos de confiança aproximados para θ em forma explícita usando as estatísticas $r_w = \text{sinal}(\hat{\theta} - \theta)w^{1/2}$ e $r_{S_R} = U(\theta)/\sqrt{K(\theta)}$ quando elas forem funções monotônicas de θ. Caso contrário, o intervalo para θ só poderá ser construído numericamente.

4.4 Inferência com Parâmetros de Incômodo

Apresenta-se aqui a teoria assintótica de primeira ordem quando o modelo estatístico contém parâmetros de perturbação. Suponha que o vetor θ de parâmetros de dimensão p é particionado como $\theta = (\psi^T, \lambda^T)^T$, onde $\dim(\psi) = q$ e $\dim(\lambda) = p - q$. Deseja-se testar $H : \psi = \psi^{(0)}$ versus $A : \psi \neq \psi^{(0)}$, onde ψ é o vetor de parâmetros de interesse e λ o vetor de parâmetros de perturbação. Seja $\ell(\psi, \lambda)$ a log-verossimilhança para ψ e λ. De agora por diante, os símbolos $\wedge$ e $\sim$ indicam quantidades estimadas segundo A e H, i.e., avaliadas nas EMV irrestrita $\hat{\theta} = (\hat{\psi}^T, \hat{\lambda}^T)^T$ e restrita $\tilde{\theta} = (\psi^{(0)T}, \tilde{\lambda}^T)^T$, respectivamente. Particionam-se o vetor escore U, a matriz de informação K e sua inversa K^{-1} da mesma maneira que θ, ou seja, $U^T = (U_\psi^T, U_\lambda^T)$,

$$K = \begin{pmatrix} K_{\psi\psi} & K_{\psi\lambda} \\ K_{\lambda\psi} & K_{\lambda\lambda} \end{pmatrix} \quad \text{e} \quad K^{-1} = \begin{pmatrix} K^{\psi\psi} & K^{\psi\lambda} \\ K^{\lambda\psi} & K^{\lambda\lambda} \end{pmatrix}.$$

Usa-se notação similar para a matriz de informação observada J e para sua inversa J^{-1}. Em geral, as quantidades U_ψ, U_λ, $K_{\psi\psi}$, $K_{\psi\lambda} = K_{\lambda\psi}^T$ e $K_{\lambda\lambda}$ dependem de ambos vetores ψ e λ.

A estatística escore baseia-se na normalidade assintótica da componente da função escore $U_\psi = U_\psi(\psi^{(0)}, \lambda)$ correspondente ao vetor de parâmetros de interesse, ou seja, no resultado

$$U_\psi \xrightarrow{\mathcal{D}} N_q(0, K^{\psi\psi^{-1}}), \tag{4.20}$$

onde $K^{\psi\psi} = K^{\psi\psi}(\psi^{(0)}, \lambda)$ é a matriz de covariância assintótica de $\hat{\psi}$. Então, a estatística escore é definida pela forma quadrática

$$S_R = \tilde{U}_\psi^T \tilde{K}^{\psi\psi} \tilde{U}_\psi, \tag{4.21}$$

onde $\tilde{U} = U_\psi(\psi^{(0)}, \tilde{\lambda})$ e $\tilde{K}^{\psi\psi} = K^{\psi\psi}(\psi^{(0)}, \tilde{\lambda})$. A vantagem da estatística escore é que ela só depende da EMV segundo a hipótese nula. A distribuição assintótica de S_R segundo $H : \psi = \psi^{(0)}$ segue diretamente de (4.20) que implica $S_R \xrightarrow{\mathcal{D}} \chi_q^2$.

O desenvolvimento da estatística de Wald é similar ao da estatística escore e decorre da normalidade assintótica da EMV $\hat{\psi}$. Como $\hat{\theta}$ tem distribuição normal p-dimensional

assintótica com matriz de covariância K^{-1}, então, $\hat{\psi}$ tem também segundo H, distribuição normal q-dimensional assintótica com média $\psi^{(0)}$ e matriz de covariância $K^{\psi\psi}$, ou seja, $\hat{\psi} - \psi^{(0)} \xrightarrow{\mathcal{D}} N_q(0, K^{\psi\psi})$. A matriz $K^{\psi\psi}$ pode ser consistentemente estimada por $K^{\psi\psi}(\hat{\psi}, \hat{\lambda})$, $K^{\psi\psi}(\psi^{(0)}, \tilde{\lambda})$, $J^{\lambda\lambda}(\hat{\psi}, \hat{\lambda})$ ou $J^{\lambda\lambda}(\psi^{(0)}, \tilde{\lambda})$. Escolhendo a primeira forma a estatística de Wald é dada por

$$W = (\hat{\psi} - \psi^{(0)})^T \hat{K}^{\psi\psi^{-1}} (\hat{\psi} - \psi^{(0)}), \tag{4.22}$$

onde $\hat{K}^{\psi\psi} = K^{\psi\psi}(\hat{\psi}, \hat{\lambda})$. Usando-se as outras matrizes de peso obtêm-se estatísticas que são assintoticamente equivalentes a (4.22). Em qualquer caso, W é uma forma quadrática correspondente à distribuição normal assintótica $N_q(0, K^{\psi\psi})$ de $\hat{\psi} - \psi^{(0)}$ e, portanto, $W \xrightarrow{\mathcal{D}} \chi^2_q$, supondo a hipótese nula verdadeira.

A razão de verossimilhança para testar $H : \psi = \psi^{(0)}$ é definida como

$$w = 2\{\ell(\psi^{(0)}, \hat{\lambda}) - \ell(\psi^{(0)}, \tilde{\lambda})\} . \tag{4.23}$$

O inconveniente de (4.23) é que w requer duas maximizações. Pode-se mostrar que $w \xrightarrow{\mathcal{D}} \chi^2_q$ segundo H (Wilks, 1938). Logo, as estatísticas (4.21) – (4.23) são equivalentes até primeira ordem, pois todas convergem sob a hipótese nula para a distribuição χ^2_q. Apresenta-se, resumidamente, a estratégia de demonstração da equivalência assintótica das estatisticas S_R, W e w. Em primeiro lugar, a fórmula da inversa de uma matriz particionada produz

$$K^{\psi\psi} = (K_{\psi\psi} - K_{\psi\lambda} K_{\lambda\lambda}^{-1} K_{\lambda\psi})^{-1},$$
$$K^{\psi\lambda} = K^{\lambda\psi^T} = -K^{\psi\psi} K_{\psi\lambda} K_{\lambda\lambda}^{-1} \text{ e}$$
$$K^{\lambda\lambda} = K_{\lambda\lambda}^{-1} - K_{\lambda\lambda}^{-1} K_{\lambda\psi} K^{\psi\lambda}.$$

Além disso, $K^{\lambda\lambda} = K_{\lambda\lambda}^{-1} + K^{\lambda\psi} K_{\psi\psi}^{-1} K^{\psi\lambda}$. A relação entre as estimativas $\tilde{\lambda}$ e $\hat{\lambda}$ é

$$\tilde{\lambda} = \hat{\lambda} + K_{\lambda\lambda}^{-1} K_{\lambda\psi} (\hat{\psi} - \psi) + O_p(n^{-1}) .$$

Recorrendo à aproximação (1.15) tem-se até primeira ordem

$$\begin{pmatrix} \hat{\psi} - \psi^{(0)} \\ \hat{\lambda} - \lambda \end{pmatrix} = \begin{pmatrix} K^{\psi\psi} & K^{\psi\lambda} \\ K^{\lambda\psi} & K^{\lambda\lambda} \end{pmatrix} \begin{pmatrix} U_\psi \\ U_\lambda \end{pmatrix}$$

que substituída em (4.22) implica, ignorando quantidades de ordem $o_p(1)$,

$$W = (K^{\psi\psi}U_\psi + K^{\psi\lambda}U_\lambda)^T K^{\psi\psi -1}(K^{\psi\psi}U_\psi + K^{\psi\lambda}U_\lambda). \tag{4.24}$$

Até primeira ordem tem-se

$$\begin{aligned} U_\psi(\psi^{(0)}, \tilde{\lambda}) &= U_\psi + \frac{\partial U_\psi}{\partial \lambda}(\tilde{\lambda} - \lambda) \\ &= U_\psi - K_{\psi\lambda}(\tilde{\lambda} - \lambda). \end{aligned}$$

Como $\tilde{\lambda} - \lambda = K^{\lambda\lambda}U_\lambda + o_p(n^{-1/2})$, vem

$$U_\psi(\psi^{(0)}, \tilde{\lambda}) = U_\psi - K_{\psi\lambda}K^{\lambda\lambda}U_\lambda \ .$$

Substituindo em (4.21) resulta até primeira ordem

$$S_R = (U_\psi - K_{\psi\lambda}K^{\lambda\lambda}U_\lambda)^T K^{\psi\psi}(U_\psi - K_{\psi\lambda}K^{\lambda\lambda}U_\lambda). \tag{4.25}$$

A razão de verossimilhança pode ser decomposta como

$$w = 2\{\ell(\hat{\psi}, \hat{\lambda}) - \ell(\psi^{(0)}, \lambda)\} - 2\{\ell(\psi^{(0)}, \tilde{\lambda}) - \ell(\psi^{(0)}, \lambda)\},$$

isto é, dada pela diferença entre duas estatísticas para testar hipóteses sem parâmetros de incômodo. Assim, usando o resultado (4.18) tem-se até primeira ordem

$$w = (U_\psi^T U_\lambda^T)\begin{pmatrix} K^{\psi\psi} & K^{\psi\lambda} \\ K^{\lambda\psi} & K^{\lambda\lambda} \end{pmatrix}\begin{pmatrix} U_\psi \\ U_\lambda \end{pmatrix} - U_\psi^T K_{\lambda\lambda}^{-1} U_\psi. \tag{4.26}$$

Com uma longa álgebra envolvendo as matrizes particionadas anteriores demonstra-se que as expressões (4.24) - (4.26) são iguais e, portanto, estabelece-se a equivalência de primeira ordem das estatísticas S_R, W e w.

Um problema que surge na realização dos testes de hipóteses e na construção de regiões de confiança para ψ é escolher dentre as estatísticas (4.21) - (4.23) que, segundo $H : \psi = \psi^{(0)}$, são assintoticamente equivalentes à distribuição χ_q^2. Claramente, a escolha

pode ser pautada no grau de dificuldade de cálculo das EMV irrestrita e restrita, notando ainda que w e S_R são invariantes em relação à reparametrização da distribuição dos dados mas a estatística de Wald não é invariante. Estas estatísticas são quantidades pivotais assintóticas para o parâmetro ψ e, portanto, podem ser usadas para construir regiões assintóticas de $100(1-\alpha)\%$ de confiança para ψ. Estas regiões são definidas por um conjunto aleatório $R(\psi) \subset \mathbb{R}^q$ dependente de y e de α tal que $P(\psi \in R(\psi)) = 1-\alpha$. Assim, regiões de $100(1-\alpha)\%$ de confiança em $\mathbb{R}^q$ para ψ são deduzidas diretamente das estatísticas escore S_R em (4.21), Wald em (4.22) e razão de verossimilhança em (4.23), produzindo

$$R_1(\psi) = \{\psi;\ \tilde{U}_\psi^T \tilde{K}^{\psi\psi} \tilde{U}_\psi \leq \chi_q^2(\alpha)\},$$
$$R_2(\psi) = \{\psi;\ (\psi-\hat{\psi})^T \hat{K}^{\psi\psi^{-1}} (\psi-\hat{\psi}) \leq \chi_q^2(\alpha)\},$$
$$R_3(\psi) = \{\psi;\ \ell(\psi,\tilde{\lambda}) \geq \ell(\hat{\psi},\hat{\lambda}) - \tfrac{1}{2}\chi_q^2(\alpha)\},$$

respectivamente. Claro que $R_3(\psi)$ é mais fácil de ser construída do que as regiões $R_1(\psi)$ e $R_2(\psi)$, pois estas últimas dependem de formas quadráticas. Observe-se que $R_3(\psi)$ é decorrente da razão de verossimilhança perfilada (vide Seção 4.5), pois $\tilde{\lambda}$ é a EMV de λ condicional a ψ. As regiões $R_1(\psi)$, $R_2(\psi)$ e $R_3(\psi)$ são assintoticamente equivalentes mas em pequenas amostras são diferentes e podem ser imprecisas. Elas são aplicáveis em qualquer problema regular na construção de regiões de confiança.

No caso do parâmetro de interesse ψ ser escalar, pode-se também construir intervalos de confiança aproximados para ψ generalizando as estatísticas sinalizadas em (4.19). Assim, obtêm-se quantidades pivotais assintóticas para ψ análogas àquelas em (4.19) dadas por

$$\begin{aligned} r_w &= \text{sinal}(\hat{\psi}-\psi) w^{1/2}, \quad r_{S_R} = \tilde{U}_\psi \tilde{K}^{\psi\psi^{1/2}}, \\ r_W &= (\hat{\psi}-\psi)/\hat{K}^{\psi\psi^{1/2}}, \quad r'_{S_R} = \tilde{U}_\psi \tilde{J}^{\psi\psi^{1/2}}, \\ r'_W &= (\hat{\psi}-\psi)/\hat{J}^{\psi\psi^{1/2}}. \end{aligned} \tag{4.27}$$

Todas as estatísticas em (4.27) têm distribuição normal $N(0,1)$ assintótica. Como $r_{S_R}^2 = S_R$, $r_W^2 = W$ e $r_w^2 = w$, os intervalos obtidos das regiões $R_1(\psi)$, $R_2(\psi)$ e $R_3(\psi)$ são idênticos àqueles baseados em r_{S_R}, r_W e r_w, respectivamente. As estatísticas r'_{S_R} e r'_W são versões assintoticamente equivalentes a r_{S_R} e r_W com informação observada no lugar de informação esperada.

Uma grande simplificação ocorre no cálculo das estatísticas w, S_R e W quando os

vetores de parâmetros ψ e λ são *ortogonais*. Neste caso, a matriz de informação é bloco-diagonal pois as matrizes $K_{\psi\lambda}$ e $K_{\lambda\psi}$ se anulam e as equações de máxima verossimilhança para determinar $\hat{\psi}$ e $\hat{\lambda}$ são separáveis. Observe-se que as expressões (4.24) – (4.26) reduzem-se, sem cálculos adicionais, a $U_{\psi}^{T} K^{\psi\psi} U_{\psi}$. Como conseqüência da informação ser bloco-diagonal, as EMV $\hat{\psi}$ e $\hat{\lambda}$ são assintoticamente independentes e a covariância assintótica de $\hat{\psi}$ quando λ é desconhecido é a mesma daquela covariância quando λ é conhecido. Um outro aspecto importante é que a EMV $\hat{\lambda}_{\psi}$ de λ condicional a ψ especificado varia pouco com ψ na vizinhança de $\hat{\psi}$, com uma variação também reduzida da EMV $\hat{\psi}_{\lambda}$ de ψ com λ especificado. Mais precisamente, se $\hat{\psi} - \psi = O_p(n^{-1/2})$, então $\hat{\lambda} - \hat{\lambda}_{\psi} = O_p(n^{-1})$. Quando não há ortogonalidade, $\hat{\lambda} - \hat{\lambda}_{\psi} = O_p(n^{-1/2})$.

Exemplo 4.4 *Suponha que uma variável aleatória Y tem função densidade dependendo de dois parâmetros μ (média) e ϕ (precisão) escrita convenientemente como*

$$f(y;\theta,\phi) = \exp[\phi\{y\theta - b(\theta)\} + \phi c_1(y) + d(\phi) + d_1(y)], \tag{4.28}$$

onde $\theta = q(\mu)$ é uma função unívoca do parâmetro μ. As funções densidade das distribuições normal $N(\mu,\phi^{-1})$, gama $G(\mu,\phi)$ e normal inversa $N^{-}(\mu,\phi)$ podem ser escritas na forma (4.28). Das condições (1.9) e (1.10) verifica-se que $E(Y) = \mu = db(\theta)/d\theta$ e $\mathrm{Var}(Y) = \phi^{-1} d^2 b(\theta)/d\theta^2$. Representa-se a função de variância de Y por $V = V(\mu) = d^2 b(\theta)/d\theta^2$ pois só depende do parâmetro θ e, portanto, somente de μ. Note-se que ϕ realmente mede a precisão de Y sendo ϕ^{-1} uma medida de dispersão. Sejam n observações iid do modelo (4.28). Apresentam-se agora as estatísticas (4.21) – (4.23) para testar a média $H_1 : \mu = \mu^{(0)}$ (versus $A_1 : \mu_1 \neq \mu^{(0)}$) com o parâmetro de precisão ϕ desconhecido, e para testar o parâmetro de precisão $H_2 : \phi = \phi^{(0)}$ (versus $A_2 : \phi \neq \phi^{(0)}$) com a média μ desconhecida. A log-verossimilhança como função de μ e ϕ é expressa por

$$\ell(\mu,\phi) = n\phi\{\overline{y}q(\mu) - b(q(\mu))\} + \phi\Sigma c_1(y_i) + nd(\phi) + \Sigma d_1(y_i)\ .$$

As componentes da função escore com relação a μ e ϕ são $U_{\mu} = \frac{n\phi}{V}(\overline{y} - \mu)$ e $U_{\phi} = n\{\overline{y}q(\mu) - b(q(\mu))\} + \Sigma c_1(y_i) + nd'(\phi)$. As EMV irrestritas são: $\hat{\mu} = \overline{y}$ e $\hat{\phi}$, obtida de

$$d'(\hat{\phi}) + \overline{y}q(\overline{y}) - b(q(\overline{y})) + \frac{1}{n}\sum_{i=1}^{n} c_1(y_i) = 0. \tag{4.29}$$

A equação (4.29) pode ser não-linear (caso da distribuição gama) ou ter soluções fechadas (casos das distribuições normal e normal inversa).

No teste de H_1 versus A_1, a EMV $\hat{\phi}$ segundo A_1 é obtida de (4.29). A EMV $\tilde{\phi}$ segundo H_1 é calculada também desta equação com $\mu^{(0)}$ no lugar de $\overline{y}$. Os parâmetros μ e ϕ são ortogonais, o que facilita o cálculo das estatísticas escore e de Wald. A informação para μ, ϕ é bloco-diagonal sendo dada por $K = \text{diag}\{\frac{n\phi}{V}, -nd''(\phi)\}$. Então, as estatísticas S_{R_1} e W_1 seguem diretamente de (4.21) - (4.22) como $S_{R_1} = \frac{n\tilde{\phi}}{\tilde{V}}(\overline{y}-\mu^{(0)})^2$ e $W_1 = \frac{n\hat{\phi}}{\hat{V}}(\overline{y}-\mu^{(0)})^2$, onde $\tilde{V} = V(\mu^{(0)})$ e $\hat{V} = V(\overline{y})$. Assim, as formas de S_{R_1} e W_1 são similares; a diferença é que as quantidades da primeira estão avaliadas em H_1 e as da segunda em A_1. A razão de verossimilhança w_1 pode ser calculada de (4.23) numa forma muito simples (Cordeiro, 1987) levando-se em consideração as equações que determinam $\hat{\phi}$ e $\tilde{\phi}$. Tem-se $w_1 = 2n\{v(\tilde{\phi}) - v(\hat{\phi})\}$, onde $v(\phi) = \phi d'(\phi) - d(\phi)$. As três estatísticas S_{R_1}, W_1 e w_1 convergem assintoticamente, quando a hipótese H_1 é verdadeira, para a distribuição χ^2_1.

No teste de $H_2 : \phi = \phi^{(0)}$ versus $A_2 : \phi \neq \phi^{(0)}$ observe-se que a EMV de μ é igual a média amostral $\overline{y}$ segundo ambas as hipóteses. Levando-se em consideração a equação (4.29) que determina $\hat{\phi}$ é fácil mostrar que a razão de verossimilhança para testar H_2 reduz-se a $w_2 = 2n\{d(\hat{\phi}) - d(\phi^{(0)}) - (\hat{\phi} - \phi^{(0)})d'(\hat{\phi})\}$. Usando-se ainda (4.29), a função escore relativa a ϕ avaliada em H_2 iguala $\tilde{U}_\phi = n\{d'(\phi^{(0)}) - d'(\hat{\phi})\}$ e, portanto, obtém-se a estatística escore $S_{R_2} = -n\{d'(\phi^{(0)}) - d'(\hat{\phi})\}^2/d''(\phi^{(0)})$. A estatística de Wald é simplesmente $W_2 = -n(\hat{\phi} - \phi^{(0)})^2 d''(\hat{\phi})$. As três estatísticas w_2, S_{R_2} e W_2 são assintoticamente equivalentes, quando H_2 é verdadeira, à distribuição χ^2_1. As formas das três estatísticas nos testes de H_1 e H_2, relativas às distribuições normal, gama e normal inversa, são facilmente obtidas destas expressões gerais a partir das funções V e $d(\phi)$ (vide exercício 1 da Seção 4.6).

Exemplo 4.5 *Considere a distribuição multinomial $Y \sim M(n, \pi)$, apresentada no exemplo 4.1, sendo o vetor π de probabilidades de dimensão p. O interesse reside em testar a hipótese que o vetor π depende de um vetor θ desconhecido de dimensão q muito menor que p, i.e., testar $H : \pi \neq \pi(\theta)$ versus $A : \pi \neq \pi(\theta)$. Seja $y = (y_1, \ldots, y_p)^T$ o vetor das freqüências observadas. Cox e Hinkley (1979, Seção 9.3) demonstram que as três estatísticas clássicas para testar H versus A têm as seguintes expressões:*

$$w = 2\sum_{i=1}^{p} y_i \log\left\{\frac{y_i}{n\pi_i(\hat{\theta})}\right\}, \quad S_R = \sum_{i=1}^{p} \frac{\{y_i - n\pi_i(\hat{\theta})\}^2}{n\pi_i(\hat{\theta})}$$

e

$$W = \sum_{i=1}^{p} \{y_i - n\pi_i(\hat{\theta})\}^2 / y_i,$$

onde $\hat{\theta}$ é a EMV de θ segundo H. Admite-se aqui que $y_i > 0$ para $i = 1, \ldots, p$. A idéia da demonstração é transformar a hipótese $H : \pi = \pi(\theta)$ na forma canônica $H : \psi = \psi^{(0)}$, λ desconhecido, usada nesta seção. Assim, as três estatísticas têm formas semelhantes àquelas expressões do exemplo 4.1 relativas ao teste de uma hipótese simples sobre π. Segundo H, elas têm assintoticamente distribuição χ^2_{p-1-q}, e o teste é conduzido comparando-se os seus valores com os pontos críticos desta distribuição.

4.5 Verossimilhança Perfilada

No caso de modelos com parâmetros de perturbação costuma-se fazer inferência usando a *verossimilhança perfilada*. Como na Seção 4.3, seja $\theta = (\psi^T, \lambda^T)^T$ o vetor de parâmetros particionado nos vetores ψ e λ de parâmetros de interesse e de incômodo, respectivamente. Seja $L(\psi, \lambda)$ a verossimilhança para ψ e λ. Denota-se por $\hat{\lambda}_\psi$ a EMV de λ para dado valor de ψ. A verossimilhança perfilada para ψ é definida por

$$\tilde{L}(\psi) = L(\psi, \hat{\lambda}_\psi) \tag{4.30}$$

e é usada em vários aspectos de forma análoga a uma verossimilhança genuína. A log-verossimilhança perfilada é $\tilde{\ell}(\psi) = \log \tilde{L}(\psi)$. A forma (4.30) sugere um procedimento de maximização em duas etapas. A primeira etapa consiste em achar o valor único $\hat{\lambda}_\psi$ que maximiza $\ell(\psi, \lambda) = \log L(\psi, \lambda)$ com respeito a λ supondo ψ fixo. A segunda etapa visa a encontrar o valor de ψ que maximiza $\tilde{\ell}(\psi)$. Em geral, $\hat{\lambda}_\psi$ difere de $\hat{\lambda}$ (EMV usual de λ) por termos de ordem $O_p(n^{-1/2})$. Os máximos de $\tilde{\ell}(\psi)$ e $\ell(\psi, \lambda)$ coincidem e, então, supondo que $\hat{\psi}$ maximiza (4.30) tem-se: $\tilde{\ell}(\hat{\psi}) \geq l(\psi)$ ou $\ell(\hat{\psi}, \hat{\lambda}_{\hat{\psi}}) \geq \ell(\psi, \hat{\lambda}_\psi) \geq \ell(\psi, \lambda)$. Assim, as EMV perfiladas $\hat{\psi}$ e $\hat{\lambda}_{\hat{\psi}}$ são iguais às EMV usuais de ψ e λ. Convém ressaltar as seguintes propriedades:

1. Se $\ell = \ell(\psi, \lambda)$ é diferenciável, $\hat{\psi}$ e $\hat{\lambda}$ são soluções das equações de máxima verossimilhança $\partial\ell/\partial\psi|_{\hat{\psi},\hat{\lambda}} = 0$, $\partial\ell/\partial\lambda|_{\hat{\psi},\hat{\lambda}} = 0$ e, para todo ψ fixado, $\hat{\lambda}_\psi$ é solução de $\partial\ell/\partial\lambda|_{\hat{\lambda}_\psi} = 0$, então a EMV $\hat{\psi}$ pode ser obtida diretamente da equação de

máxima verossimilhança perfilada com o vetor λ efetivamente eliminado, i.e., de $\partial\tilde{\ell}(\psi)/\partial\psi|_{\hat{\psi}} = 0$.

2. A razão de verossimilhança perfilada $\tilde{w} = 2\{\tilde{\ell}(\hat{\psi}) - \tilde{\ell}(\psi^{(0)})\}$ é igual à razão de verossimilhança usual para testar a hipótese $H : \psi = \psi^{(0)}$, i.e.,

$$\tilde{w} = 2\{\tilde{\ell}(\hat{\psi}) - \tilde{\ell}(\psi^{(0)})\} = 2\{\ell(\hat{\psi}, \hat{\lambda}) - \ell(\psi^{(0)}, \hat{\lambda})\}$$

e, portanto, $\tilde{w} \xrightarrow{\mathcal{D}} \chi^2_q$, onde $q = \dim(\psi)$.

3. A região de confiança perfilada

$$R(\psi) = \{\psi;\ \tilde{\ell}(\hat{\psi}) - \tilde{\ell}(\psi) \leq c\}$$

é uma região de confiança aproximada para ψ com o nível de significância determinado da distribuição χ^2_q assintótica de $\tilde{w}$. Quando $q \leq 5$, $c = q + 1$, $q + 3$ e $q + 5$ produz regiões de 95%, 99% e 99,9% de confiança para ψ.

4. A inversa da informação observada perfilada $\tilde{J}(\psi)$ para ψ é simplesmente dada por

$$\tilde{J}(\psi)^{-1} = J^{\psi\psi}(\psi, \hat{\lambda}_\psi),$$

ou seja, é igual ao bloco (ψ, ψ) da inversa da matriz de informação observada usual $J(\psi, \lambda)$ avaliada em $(\psi^T, \hat{\lambda}_\psi^T)^T$. A estrutura de covariância assintótica de $\hat{\psi}$ pode ser estimada por $\tilde{J}(\hat{\psi})^{-1}$.

4.6 Exercícios

1. Nas distribuições normal $N(\mu, \phi^{-1})$, gama $G(\mu, \phi)$ e normal inversa $N^-(\mu, \phi)$ do exemplo 4.4 apresente as formas das estatísticas w, S_R e W para os testes da média μ e do parâmetro de precisão ϕ. Obtenha regiões de confiança baseadas nestas três estatísticas para μ (ϕ desconhecido) e ϕ (μ desconhecido).

2. Suponha que se deseja construir intervalos de confiança para μ na distribuição normal $N(\mu, 1)$. Compare os intervalos de confiança para μ baseados nas estatísticas r_w, r_{S_R} e r_W dadas em (4.19) com o intervalo exato de $100(1 - \alpha)\%$ de confiança para μ. Ilustre numericamente a comparação.

3. Calcule as estatísticas w, S_R e W para testar o parâmetro θ nas seguintes distribuições uniparamétricas: Cauchy $CA(\theta)$, série logaritmica $SL(\theta)$ e série de potências $SP(\theta)$.

4. Suponha a distribuição de Weibull do exemplo 1.5. Obtenha as formas das estatísticas escore, razão de verossimilhança e Wald para testar α (ϕ desconhecido) e ϕ (α desconhecido).

5. Deduza a melhor região crítica para testar $H : \theta = \theta^{(0)}$ versus $A : \theta = \theta^{(1)}$ supondo que a função modelo é $f(y;\theta) = c(\theta)d(y)\exp\{a(\theta)b(y)\}$.

6. Calcular a MRC para testar uma hipótese simples $H : \mu = \mu^{(0)}$ versus uma alternativa simples $A : \mu = \mu^{(1)}$ nos casos de μ ser a média da distribuição de Poisson $P(\mu)$ e ser a probabilidade de sucesso na distribuição binomial $B(m,\mu)$.

7. Suponha a família de locação e escala definida por

$$f(y;\mu,\sigma) = \sigma^{-1} f(\frac{y-\mu}{\sigma}),$$

onde $\mu \in \mathbb{R}$, $y \in \mathbb{R}$ e $\sigma > 0$. Deduza as formas das estatísticas w, W e S_R para testar as hipóteses $H_1 : \mu = \mu^{(0)}$ com σ desconhecido e $H_2 : \sigma = \sigma^{(0)}$ com μ desconhecido.

8. Sejam $f_0(\cdot)$ e $f_1(\cdot)$ duas funções densidades com o mesmo suporte. Forma-se a família de densidades

$$f_\psi(y) = c(\psi) f_0(y)^{1-\psi} f_1(y)^\psi,$$

onde $c(\psi)$ é uma função normalizadora. Desenvolva uma estatística escore para testar a hipótese $H : \psi = 0$ baseada em n observações *iid* de $f_\psi(y)$.

9. Nas distribuições normal $N(\mu,\phi)$, normal inversa $N^-(\mu,\phi)$ e gama $G(\mu,\phi)$, obtenha regiões aproximadas de $100(1-\alpha)\%$ de confiança para: (a) μ quando ϕ é desconhecido; (b) para ϕ quando μ é desconhecido.

10. Sejam $y_1,\ldots,y_n$ observações de Poisson com médias $\mu_1,\ldots,\mu_n$ dadas por $\log \mu_i = \alpha + \beta x_i$, onde $x_1,\ldots,x_n$ são valores de uma covariável x conhecida. Determine intervalos de confiança aproximados para β baseados nas estatísticas escore, Wald e da razão de verossimilhança.

Capítulo 5

Teoria Assintótica de Segunda Ordem

5.1 Introdução

Neste capítulo apresentam-se alguns resultados referentes à teoria assintótica de segunda ordem, que são refinamentos dos resultados gerais do Capítulo 4. Agora, os erros associados às propriedades estatísticas são em geral de ordem $O(n^{-2})$ ao invés de ordem $O(n^{-1})$, como na teoria assintótica de primeira ordem. As pesquisas em teoria assintótica de segunda ordem têm crescido a passos largos nos últimos anos, principalmente em relação aos seguintes tópicos: correção do viés da EMV, fórmula aproximada de Barndorff-Nielsen para a função densidade da EMV, cálculo aproximado da função de distribuição da EMV, correções de Bartlett para melhorar os testes baseados na razão de verossimilhança e extensão para as correções tipo-Bartlett de outras estatísticas de teste. Neste capítulo, apresenta-se um estudo resumido de cada um destes tópicos, citando-se algumas das principais referências para estudos posteriores.

5.2 Identidades de Bartlett

Seja $L = L(\theta)$ a verossimilhança total de um problema regular supondo que as observações são independentes mas não necessariamente identicamente distribuídas, onde θ é um vetor de $\mathbb{R}^p$. Adota-se a seguinte notação para as derivadas da log-verossimilhança

$\ell = \ell(\theta) = \log L(\theta)$, onde todos os índices variam de 1 a p : $U_r = \partial\ell/\partial\theta_r$, $U_{rs} = \partial^2\ell/\partial\theta_r\partial\theta_s$, etc. Os *momentos conjuntos* de derivadas de $\ell(\theta)$ são $\mu_r = E(U_r)$, $\mu_{rs} = E(U_{rs})$, $\mu_{r,s} = E(U_rU_s)$, $\mu_{r,st} = E(U_rU_{st})$ e assim por diante. Como $\mu_r = 0$, os correspondentes *cumulantes conjuntos* ($\kappa's$) expressos em termos dos momentos são: $\kappa_{r,s} = \mu_{r,s}$, $\kappa_{rs} = \mu_{rs}$, $\kappa_{rs,t} = \mu_{rs,t}$, $\kappa_{rs,tu} = \mu_{rs,tu} - \mu_{rs}\mu_{tu}$, $\kappa_{r,s,t} = \mu_{r,s,t}$ e $\kappa_{r,s,t,u} = \mu_{r,s,t,u} - \Sigma_{(3)}\mu_{r,s}\mu_{t,u}$, onde $\Sigma_{(k)}$ representa o somatório sobre todas as k combinações de índices, etc. Os momentos e cumulantes acima não são independentes, mas satisfazem certas equações que facilitam seus cálculos. Estas equações, que representam condições de regularidade, são denominadas de *identidades de Bartlett*. As mais importantes são: $\kappa_r = 0$ e $\kappa_{rs} + \kappa_{r,s} = 0$. Os cumulantes $\kappa's$ referem-se a um total sobre a amostra e, em geral, são da ordem $O(n)$. A idéia central na dedução das identidades de Bartlett é a validade em problemas regulares da fórmula $\frac{\partial}{\partial\theta}E\{t(Y)\} = \int t(y)\frac{\partial f(y;\theta)}{\partial\theta}dy$ para qualquer estatística $t(Y)$, ou seja, pode-se inverter a ordem das operações de diferenciação em relação a θ e integração com respeito a y. Mostra-se, nesta seção, como obter algumas identidades de Bartlett. As outras identidades poderão ser deduzidas de forma semelhante por diferenciações sucessivas em relação às componentes de θ. Expressando as identidades em termos dos cumulantes, outras identidades análogas podem ser deduzidas para os momentos.

As derivadas de cumulantes são escritas com sobrescritos: $\kappa_{rs}^{(t)} = \partial\kappa_{rs}/\partial\theta_t$, $\kappa_{rs}^{(tu)} = \partial^2\kappa_{rs}/\partial\theta_t\partial\theta_u$, $\kappa_{rst}^{(u)} = \partial\kappa_{rst}/\partial\theta_u$, etc. Da definição da função escore tem-se $U_r = L_r/L$, onde $L_r = \partial L/\partial\theta_r$. Diferenciando $\int L dy = 1$ em relação a θ_r vem $\int L_r dy = 0$ e, então, $\kappa_r = E(U_r) = 0$. Diferenciando a última integral em relação a θ_s, encontra-se $\int(U_{rs}L + U_rU_sL)dy = 0$, ou seja, $\kappa_{rs} + \kappa_{r,s} = 0$. Diferenciando novamente a integral em relação a θ_t obtém-se $\kappa_{r,s,t} + \kappa_{rst} + \Sigma_{(3)}\kappa_{r,st} = 0$. Outras identidades de Bartlett são deduzidas de forma análoga:

$$\begin{aligned}
\kappa_{r,st} + \kappa_{rst} - \kappa_{st}^{(r)} &= 0,\ \kappa_{r,s,t} - 2\kappa_{rst} + \Sigma_{(3)}\kappa_{rs}^{(t)} = 0,\\
\kappa_{rst}^{(u)} &= \kappa_{rst,u},\ \kappa_{r,stu} + \kappa_{rstu} - \kappa_{stu}^{(r)} = 0,\\
\kappa_{rstu} + \Sigma_{(4)}\kappa_{r,stu} + \Sigma_{(3)}\kappa_{rs,tu} + \Sigma_{(6)}\kappa_{r,s,tu} + \kappa_{r,s,t,u} &= 0,\\
\kappa_{r,s,t,u} &= -3\kappa_{rstu} + 2\Sigma_{(4)}\kappa_{rst}^{(u)} - \Sigma_{(6)}\kappa_{rs}^{(tu)} + \Sigma_{(3)}\kappa_{rs,tu},\\
\kappa_{r,s,tu} &= \kappa_{rstu} - \kappa_{rtu}^{(s)} - \kappa_{stu}^{(r)} + \kappa_{tu}^{(rs)} - \kappa_{rs,tu},\ \text{etc.}
\end{aligned}$$

Claro que no caso uniparamétrico basta coincidir os índices para encontrar $\kappa_{\theta,\theta\theta} + \kappa_{\theta\theta\theta} - \kappa_{\theta\theta}^{(\theta)} = 0$, $\kappa_{\theta,\theta,\theta} - 2\kappa_{\theta\theta\theta} + 3\kappa_{\theta\theta}^{(\theta)} = 0$, e assim por diante.

A grande vantagem das identidades de Bartlett é facilitar a obtenção dos cumulantes $\kappa's$, pois determinada parametrização pode conduzir a um cálculo direto simples de alguns cumulantes, sendo os demais calculados indiretamente através destas identidades. Esses cumulantes têm como aplicabilidade principal o cálculo do viés de segunda ordem da EMV (Seção 5.3) e das correções de Bartlett (Seção 5.6) e tipo-Bartlett (Seção 5.7).

Exemplo 5.1 *Considere a distribuição normal* $N(\mu, \sigma^2)$ *cuja log-verossimilhança* $\ell = \ell(\theta)$ *para* $\theta = (\mu, \sigma^2)^T$, *baseada numa amostra iid de tamanho* n, *é dada por*

$$\ell = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2 \, .$$

Os cumulantes são facilmente obtidos como $\kappa_{\mu\mu} = -n/\sigma^2$, $\kappa_{\sigma^2\sigma^2} = -n/2\sigma^4$, $\kappa_{\mu\sigma^2} = 0$, $\kappa_{\mu,\mu,\mu} = \kappa_{\mu,\mu\mu} = \kappa_{\mu\mu\mu} = 0$, $\kappa_{\sigma^2,\sigma^2,\sigma^2} = -\kappa_{\sigma^2,\sigma^2\sigma^2} = n/\sigma^6$, $\kappa_{\sigma^2\sigma^2\sigma^2} = 2n/\sigma^6$, $\kappa_{\mu,\mu\sigma^2} = -\kappa_{\mu\mu\sigma^2} = -n/\sigma^4$, $\kappa_{\mu,\mu,\sigma^2} = 3n/\sigma^4$, $\kappa_{\mu\mu\sigma^2\sigma^2} = -2n/\sigma^6$, *etc., muitos deles através das identidades de Bartlett.*

5.3 Correção do Viés da EMV

As EMV são, em geral, viesadas para os valores verdadeiros dos parâmetros em modelos não-lineares quando o tamanho n dos dados é pequeno ou a informação de Fisher é reduzida. Muitas vezes o viés é ignorado na prática, justificando-se que ele é desprezível quando comparado ao erro padrão da EMV. De fato, o viés é de ordem n^{-1} enquanto que o desvio padrão da estimativa é de ordem $n^{-1/2}$. Entretanto, para alguns modelos não-lineares, o viés em pequenas amostras pode ser apreciável e ter magnitude comparável ao erro padrão da EMV. Em modelos uniparamétricos, Bartlett (1953) deduziu uma fórmula para o viés de ordem n^{-1} da EMV no caso *iid.* Nos modelos multiparamétricos, os vieses de ordem n^{-1} das EMV, supondo observações independentes mas não necessariamente identicamente distribuídas, foram deduzidos em generalidade por Cox e Snell (1968).

Considere um modelo estatístico $f(y;\theta)$ com $\theta \in \mathbb{R}^p$. Seja $\hat{\theta}$ a EMV de θ obtida como

solução do sistema de equações de máxima verossimilhança $\hat{U}_r = 0$ para $r = 1, \ldots, p$. Suponha que as condições (i) – (v) dadas na Seção 4.2 sejam satisfeitas. Expandindo $\hat{U}_r = 0$ até primeira ordem vem $U_r + \sum_s U_{rs}(\hat{\theta}_s - \theta_s) + O_p(1) = 0$ e, em notação matricial, $U = J(\hat{\theta} - \theta) + O_p(1)$. Como $J = K + O_p(n^{1/2})$ tem-se $U = K(\hat{\theta} - \theta) + O_p(1)$ e, então,

$$\hat{\theta} - \theta = K^{-1}U + O_p(n^{-1}) \, . \tag{5.1}$$

A fórmula (5.1) (idêntica a (1.14) mais o erro estocástico) desempenha um papel importante no cálculo de momentos e cumulantes da EMV de ordens superiores. Expandindo $\hat{U}_r$ até segunda ordem resulta

$$U_r + \sum_s U_{rs}(\hat{\theta}_s - \theta_s) + \frac{1}{2}\sum_{s,t} U_{rst}(\hat{\theta}_s - \theta_s)(\hat{\theta}_t - \theta_t) + o_p(1) = 0 \tag{5.2}$$

e calculando o seu valor esperado encontra-se que

$$\sum_s \kappa_{rs} E(\hat{\theta}_s - \theta_s) + \sum_s \mathrm{Cov}(U_{rs}, \hat{\theta}_s - \theta_s) + \frac{1}{2}\sum_{s,t} \kappa_{rst}(-\kappa^{st}) + o(1) = 0, \tag{5.3}$$

onde $-\kappa^{rs} = \kappa^{r,s}$ representa o elemento (r, s) da inversa K^{-1} da matriz de informação. Segue-se o cálculo de $\mathrm{Cov}(U_{rs}, \hat{\theta}_s - \theta_s)$ até $o(1)$ com o uso de (5.1): $\mathrm{Cov}(U_{rs}, \hat{\theta}_s - \theta_s) = \mathrm{Cov}\left(U_{rs}, -\sum_t \kappa^{st} U_t\right) = -\sum_t \kappa_{rs,t}\kappa^{st}$. Definindo o viés de ordem n^{-1} de $\hat{\theta}_r$ por $B(\hat{\theta}_r)$ e substituindo a última expressão em (5.3) obtém-se

$$\sum_s \kappa_{rs} B(\hat{\theta}_s) = \sum_{s,t} \kappa^{st}\left(\kappa_{rs,t} + \frac{1}{2}\kappa_{rst}\right) + o(1),$$

cuja inversão produz para $r = 1, \ldots, p$ até $O(n^{-1})$

$$B(\hat{\theta}_r) = \sum_{s,t,u} \kappa^{ru}\kappa^{st}\left(\kappa_{rs,t} + \frac{1}{2}\kappa_{rst}\right) \, . \tag{5.4}$$

A fórmula (5.4), devida a Cox e Snell (1968), é bastante geral para determinar o viés de ordem $O(n^{-1})$ da EMV em modelos multiparamétricos. Para calculá-lo basta conhecer a inversa da matriz de informação e os cumulantes $\kappa_{rs,t}$ e κ_{rst} em relação a todos os parâmetros. A expressão $(\kappa_{rs,t} + \frac{1}{2}\kappa_{rst})$ na fórmula (5.4) pode ser substituída

por $\kappa_{rs}^{(t)} - \kappa_{rst}/2$, como conseqüência da identidade de Bartlett $\kappa_{t,rs} + \kappa_{rst} - \kappa_{rs}^{(t)} = 0$. Em muitas situações multiparamétricas, torna-se conveniente colocar a equação (5.4) em notação matricial (vide Cordeiro e McCullagh, 1991 e Cordeiro e Klein, 1994).

A grande utilidade da equação (5.4) é definir uma EMV corrigida até $O(n^{-1})$ dada por $\tilde{\theta}_r = \hat{\theta} - \hat{B}(\hat{\theta}_r)$, onde $\hat{B}(\cdot)$ é o viés $B(\cdot)$ avaliado em $\hat{\theta}$. A EMV corrigida $\tilde{\theta}_r$ tem viés de ordem n^{-2}, isto é, $E(\tilde{\theta}) = \theta + O(n^{-2})$, e pode ser preferida em relação à EMV usual $\hat{\theta}_r$ cujo viés é de ordem $O(n^{-1})$. Diz-se que $\hat{\theta}$ é *a EMV de primeira ordem* (seu viés é de ordem n^{-1}) enquanto $\tilde{\theta}$ é *a EMV de segunda ordem* (seu viés é de ordem n^{-2}).

Exemplo 5.2 *Como no exemplo anterior, considere n observações iid de uma distribuição normal $N(\mu, \sigma^2)$. O interesse reside em calcular os vieses de ordem n^{-1} das EMV dos parâmetros μ (média) e σ (desvio padrão). Note-se que os cumulantes aqui são calculados em relação a (μ, σ) e não, como no exemplo 5.1, em relação a (μ, σ^2). Os elementos da matriz de informação para μ e σ seguem de imediato como $\kappa_{\mu,\mu} = n/\sigma^2$, $\kappa_{\mu,\sigma} = 0$ e $\kappa_{\sigma,\sigma} = 2n/\sigma^2$. Os cumulantes de terceira ordem são calculados sem maiores dificuldades: $\kappa_{\mu\mu\mu} = \kappa_{\mu,\mu\mu} = \kappa_{\sigma,\mu\mu} = \kappa_{\sigma,\mu\sigma} = \kappa_{\mu,\sigma\sigma} = \kappa_{\mu\sigma\sigma} = 0$, $\kappa_{\mu\mu\sigma} = -\kappa_{\mu,\mu\sigma} = 2n/\sigma^3$, $\kappa_{\sigma,\sigma\sigma} = -6n/\sigma^3$ e $\kappa_{\sigma\sigma\sigma} = 10n/\sigma^3$. Logo, usando a equação (5.4) vem $B(\hat{\mu}) = 0$, como esperado, pois $\hat{\mu} = \Sigma y_i/n$ não é viesado; com alguma álgebra, obtém-se $B(\hat{\sigma}) = -3\sigma/4n$. Este resultado está de acordo com o viés exato de $\hat{\sigma} = \{\Sigma(y_i - \overline{y})^2/n\}^{1/2}$ dado por $E(\hat{\sigma}) = \sqrt{\frac{2}{n}}\frac{\Gamma(\frac{n-1}{2})}{\Gamma(\frac{n}{2})}\sigma$, que é deduzido da distribuição χ^2_{n-1} de $(n-1)\hat{\sigma}^2/\sigma^2$. Com efeito, usando a expansão de Stirling em $E(\hat{\sigma})$ implica $E(\hat{\sigma}) = \sigma\{1 - \frac{3}{4n} + O(n^{-2})\}$. A EMV corrigida de σ é, então, $\tilde{\sigma} = (1 + \frac{3}{4n})\hat{\sigma}$.*

No caso de um modelo uniparamétrico $f(y;\theta)$ com $\theta \in \mathbb{R}$, o viés de ordem n^{-1} segue de (5.4) fazendo todos os índices iguais a θ. Tem-se,

$$B(\hat{\theta}) = \kappa^{\theta\theta^2}\left(\kappa_{\theta\theta,\theta} + \frac{1}{2}\kappa_{\theta\theta\theta}\right) = \kappa^{\theta\theta^2}\left(\kappa_{\theta\theta}^{(\theta)} - \frac{1}{2}\kappa_{\theta\theta\theta}\right). \tag{5.5}$$

Exemplo 5.3 *Considere n observações iid de uma distribuição exponencial de taxa ρ cuja função densidade é $f(y;\rho) = \rho e^{-\rho y}$. A informação para ρ é $\kappa_{\rho,\rho} = n/\rho^2$ e os cumulantes $\kappa_{\rho\rho\rho}$ e $\kappa_{\rho\rho,\rho}$ são $2n/\rho^3$ e 0, respectivamente. A EMV de ρ é dada por $\rho = 1/\overline{y}$ e seu viés de ordem n^{-1} segue de (5.5) como $B(\hat{\rho}) = \rho/n$. Logo, a EMV corrigida da taxa da*

distribuição exponencial é simplesmente $\tilde{\rho} = (1 - \frac{1}{n})\hat{\rho}$. *O viés* $B(\hat{\rho}) = \rho/n$ *pode ser obtido por primeiros princípios expandindo* $\hat{\rho}$ *em série de Taylor ao invés de usar a equação (5.5). Entretanto, na prática, o cálculo do viés a partir de (5.5) é bem mais freqüente.*

O cálculo de momentos e cumulantes da EMV $\hat{\theta}$ de ordem superior, como por exemplo $E(\hat{\theta})$ e $\mathrm{Var}(\hat{\theta})$ até $O(n^{-2})$, é bastante complicado e envolve a inversão da expansão (5.2). Esta inversão se torna bastante complexa à medida em que se incluem na equação (5.2) termos de ordem inferior. A inversão da expansão (5.2) produz, até ordem $O_p(n^{-3/2})$,

$$\begin{aligned} \hat{\theta}_r - \theta_r &= -\sum_s \kappa^{rs} U_s + \sum_{s,t,u} \kappa^{rs}\kappa^{tu}(U_{st} - \kappa_{st})U_u \\ &- \frac{1}{2}\sum_{s,t,i,j,k} \kappa^{ri}\kappa^{sj}\kappa^{tk}\kappa_{ijk}U_sU_t + O_p(n^{-3/2}) \,. \end{aligned} \tag{5.6}$$

O primeiro termo do lado direito de (5.6) é de ordem $O_p(n^{-1/2})$ e os outros dois são de ordem $O_p(n^{-1})$, sendo o erro $O_p(n^{-3/2})$. Para obter, por exemplo, $\mathrm{Var}(\hat{\theta}_r)$ até $O(n^{-2})$ ($-\kappa^{rr}$ é o seu termo de ordem n^{-1}), eleva-se (5.6) ao quadrado e calcula-se seu valor esperado até a ordem desejada. O cálculo é complicado mesmo no caso uniparamétrico. O leitor poderá consultar o livro de Shenton e Bowman (1977), que fornece em toda sua extensão detalhes destes cálculos. Em especial, estes autores apresentam fórmulas gerais para $E(\hat{\theta})$ e $\mathrm{Var}(\hat{\theta})$ até ordem n^{-2} no caso multiparamétrico, e para os quatro primeiros momentos de $\hat{\theta}$ até as ordens n^{-2}, n^{-3}, n^{-3} e n^{-4}, respectivamente, no caso uniparamétrico. Neste caso, Ferrari et al. (1996) obtiveram EMV corrigidas até segunda e terceira ordens e compararam seus erros padrão. A partir da fórmula (5.4) para o viés de ordem n^{-1}, pode-se, alternativamente, calcular o viés da EMV até ordem n^{-2} no caso multiparamétrico (caso realmente seja necessário) usando a técnica "jacknife" (Cox e Hinkley, 1979, Seção 8.4).

Na década de 90 vários artigos foram publicados apresentando expressões matriciais simples para os vieses das EMV em modelos de regressão. Estas expressões são fáceis de serem implementadas pois não dependem do cálculo dos cumulantes, sendo funções apenas das características (de cunho estatístico) do modelo. Cordeiro e McCullagh (1991) obtiveram uma fórmula matricial geral para os vieses de ordem n^{-1} das EMV nos modelos lineares generalizados. Cordeiro (1993) também obteve, em notação matricial, fórmulas

de segunda ordem das EMV em dois modelos heterocedásticos de regressão. Cordeiro e Klein (1994) deduziram fórmulas matriciais para os vieses de segunda ordem das EMV em modelos ARMA. Paula e Cordeiro (1995) obtiveram fórmulas para os vieses de ordem n^{-1} das EMV dos parâmetros em modelos não-exponenciais não-lineares. Finalmente, Cordeiro e Cribari-Neto (1998) concluíram, através de estudos de simulação dos vieses das EMV nos modelos não-exponenciais não-lineares, que as EMV corrigidas são mais precisas em termos de erro médio quadrático do que as estimativas usuais.

5.4 Função Densidade da EMV

Seja Y uma variável aleatória cuja função geratriz de cumulantes $K(t)$ é conhecida. A aproximação ponto de sela para a função densidade $f_Y(y)$ de Y é obtida da equação (3.22) fazendo $n = 1$, ou seja:

$$f_Y(y) \doteq \frac{1}{\sqrt{2\pi K''(\hat{\theta})}} \exp\{K(\hat{\theta}) - \hat{\theta}y\} , \tag{5.7}$$

onde $\hat{\theta}$ é determinado por $K'(\hat{\theta}) = y$. A generalidade da equação (5.7) permite aplicá-la para aproximar um grande número de funções densidade com o conhecimento das suas correspondentes funções geratrizes de cumulantes $K(t)$. Para isso basta resolver as equações $K'(\hat{\theta}) = y$ e calcular $\hat{\theta}$.

A função geratriz de cumulantes aparece naturalmente nos modelos exponenciais uniparamétricos dados por

$$f_Y(y;\theta) = \exp\{\theta y - b(\theta) + h(y)\}, \tag{5.8}$$

sendo trivialmente obtida como $K(t) = b(\theta+t) - b(\theta)$. A log-verossimilhança para θ dado y é $\ell(\theta;y) = \theta y - b(\theta)$ mais uma constante arbitrária que não depende de θ. Assim, a aproximação ponto de sela (5.7) para o modelo exponencial (5.8) pode ser escrita como

$$f_Y(y;\theta) \doteq \frac{1}{\sqrt{2\pi J(\hat{\theta})}} \exp\{\ell(\theta;y) - \ell(\hat{\theta};y)\}, \tag{5.9}$$

onde $\hat{\theta} = \hat{\theta}(y)$ é a EMV de θ decorrente da equação $K'(\hat{\theta}) = b'(\hat{\theta}) = y$ e $J(\hat{\theta}) = -\frac{d^2\ell(\theta;y)}{d\theta^2}|_{\theta=\hat{\theta}}$

é a informação observada avaliada em $\hat{\theta}$. A aproximação (5.9) pode agora ser transformada para obter a aproximação correspondente da função densidade de $\hat{\theta}$, implicando

$$f_{\hat{\theta}}(\theta;y)\doteq\frac{1}{\sqrt{2\pi}}J(\hat{\theta})^{1/2}\exp\{\ell(\theta;y)-\ell(\hat{\theta};y)\}\ . \tag{5.10}$$

A equação (5.10) define uma aproximação para a função densidade da EMV $\hat{\theta}$ de θ no modelo exponencial (5.8). O erro associado a (5.10) é multiplicativo da forma $1+O(n^{-3/2})$.

A equação (5.10) pode ser generalizada para o modelo exponencial (1.18) de ordem p, substituindo $\sqrt{2\pi}$ por $(2\pi)^{p/2}$ e $J(\hat{\theta})$ pelo determinante da matriz de informação observada em $\hat{\theta}$, isto é, $|J(\hat{\theta})|$, resultando em

$$f_{\hat{\theta}}(y;\theta)\doteq(2\pi)^{-p/2}|J(\hat{\theta})|^{1/2}\exp\{\ell(\theta;y)-\ell(\hat{\theta};y)\}\ . \tag{5.11}$$

Esta equação é conhecida como *aproximação de Barndorff-Nielsen* para a função densidade de $\hat{\theta}$. Ela tem propriedades interessantes como a invariância segundo transformação um-a-um dos dados e, também, segundo reparametrização, isto é, se w e θ são parametrizações alternativas então, em óbvia notação, as aproximações calculadas de (5.11) para as funções densidade de $\hat{w}$ e $\hat{\theta}$ satisfazem

$$f_{\hat{w}}(w;y)=f_{\hat{\theta}}(\theta;y)\left|\frac{\partial\hat{\theta}}{\partial\hat{w}}\right|\ . \tag{5.12}$$

A fórmula (5.11) pode incluir uma constante de proporcionalidade $c(\theta)$ visando tornar sua integral igual a um sobre o suporte de $\hat{\theta}$. Esta constante é, também, invariante segundo reparametrização. Barndorff-Nielsen (1983) examinou a validade da equação (5.11) para distribuições multiparamétricas fora da família exponencial.

Exemplo 5.4 *Suponha que n observações iid sejam obtidas da distribuição exponencial com média μ. A log-verossimilhança para μ é dada por $\ell(\mu;y)=-n\log\mu-n\hat{\mu}/\mu$, onde $\hat{\mu}=\overline{y}$ e $J(\mu)=n/\mu^2$ é a informação observada para μ. A aproximação para a função densidade de $\hat{\mu}$ segue de (5.10) como*

$$f_{\hat{\mu}}(\mu;y)\doteq\overline{\Gamma}(n)^{-1}\left(\frac{\hat{\mu}}{\mu}\right)^{n-1}\frac{1}{\mu}e^{-n\hat{\mu}/\mu}, \tag{5.13}$$

onde $\overline{\Gamma}(n) = (2\pi)^{1/2} n^{n-0,5} e^{-n}$ *é a aproximação de Stirling para* $\Gamma(n)$. *Em especial, pode-se demonstrar que normalizando (5.13) obtém-se a função densidade exata de* $\hat{\mu}$. *Se o parâmetro* $\rho = \mu^{-1}$ *é usado para especificar a distribuição exponencial, tem-se* $\hat{\rho} = \overline{y}^{-1}$ *e, com uma simples mudança de notação, vem* $\ell(\rho; y) = n \log \rho - n\rho/\hat{\rho}$ *e* $J(\rho) = n/\rho^2$. *Assim, a aproximação (5.10) para a função densidade de* $\hat{\rho}$ *fica de acordo com (5.13), ilustrando a propriedade (5.12) de invariância.*

Exemplo 5.5 *Considere a distribuição Gaussiana inversa com parâmetros* $\lambda > 0$ *e* $\alpha > 0$, *supondo* α *conhecido, cuja função densidade é dada por*

$$f_Y(y; \alpha, \lambda) = \sqrt{\frac{\lambda}{2\pi}} e^{\sqrt{\alpha\lambda}} y^{-3/2} \exp\left\{-\frac{1}{2}\left(\frac{\lambda}{y} + \alpha y\right)\right\} .$$

Considere uma amostra de n observações iid desta distribuição. Demonstra-se, usando (5.10), que a função densidade de $\hat{\lambda}$ *pode ser escrita como*

$$f_{\hat{\lambda}}(\lambda; y, \alpha) = \hat{\lambda}^{-\frac{n}{2}-1}(1 + \sqrt{\alpha\hat{\lambda}}/2)^{1/2} \exp\left\{-\frac{n}{2}\sqrt{\alpha\hat{\lambda}} - \frac{n}{2}\left(\frac{\lambda}{\hat{\lambda}}\right)(1 + \sqrt{\alpha\hat{\lambda}})\right\},$$

onde $\hat{\lambda} = 4\{(\alpha + 4n^{-1}\Sigma y_i^{-1})^{1/2} - \sqrt{\alpha}\}^{-2}$.

Exemplo 5.6 *Considere a função densidade da distribuição gama com parâmetros* μ *(média) e* ν *(índice) desconhecidos. Tem-se*

$$f_Y(y; \mu, \nu) = \left(\frac{\nu}{\mu}\right)^{\nu} y^{\nu-1} e^{-\nu y/\mu} / \Gamma(\nu) .$$

Considere n observações iid desta distribuição. A EMV de $\theta = (\mu, \nu)^T$ *é deduzida de* $\hat{\mu} = \overline{y}$ *e* $\log \hat{\nu} - \psi(\hat{\nu}) = \log(\overline{y}/\tilde{y})$, *onde* $\overline{y}$ *e* $\tilde{y}$ *são as médias aritmética e geométrica dos dados. Com alguma álgebra, demonstra-se através de (5.11) que a função densidade* $f_{\hat{\theta}}(\mu, \nu; y)$ *de* $\hat{\theta} = (\hat{\mu}, \hat{\nu})^T$ *admite a decomposição*

$$f_{\hat{\theta}}(\mu, \nu; y) = f_{1\hat{\mu}}(\mu; \nu, y) f_{2\hat{\nu}}(\nu; y),$$

onde

$$f_{1\hat{\mu}}(\mu; \nu, y) = \left(\frac{\nu}{\mu}\right)^{n\nu} \hat{\mu}^{n\nu-1} \exp(-n\nu\hat{\mu}/\mu)$$

e

$$f_{2\hat{\nu}}(\nu;y) = \{\Gamma(\hat{\nu})\Gamma(\nu)\}^n\{\hat{\nu}\psi'(\hat{\nu}) - 1\}^{1/2}\exp[n\{(\hat{\nu}-\nu)\psi(\hat{\nu}) + \hat{\nu} - \nu\log\hat{\nu}\}] .$$

A decomposição acima revela que as EMV $\hat{\mu}$ e $\hat{\nu}$ são independentes até a ordem considerada pela aproximação (5.11). Adicionalmente, a aproximação $f_{1\hat{\mu}}(\mu;\nu,y)$ para a função densidade de $\hat{\mu}$ é exata após renormalização.

Se a dimensão de y é maior do que p, então $f_{\hat{\theta}}(\theta;y)$ deve ser interpretada como a função densidade condicional $f_{\hat{\theta}|t}(\theta;y)$ de $\hat{\theta}$ dada alguma estatística $t = t(y)$, exatamente ou aproximadamente ancilar, isto é, a distribuição marginal de $t = t(y)$ não depende de θ, pelo menos aproximadamente. O leitor poderá consultar as seguintes referências para obter maiores detalhes da equação (5.11): Barndorff-Nielsen (1983, 1986, 1988), McCullagh (1984), Reid (1988), Fraser (1988, 1990) e Barndorff-Nielsen e Cox (1994, Seções 6.2 e 7.4).

5.5 Cálculo de Probabilidades Baseado na Verossimilhança

Para uma variável aleatória Y com função geratriz de cumulantes $K(t)$, a equação de Lugannani e Rice (1980) para aproximar sua função de distribuição $F_Y(y)$ é dada por

$$F_Y(y) = P(Y \leq y) \doteq \Phi(z) + \phi(z)\left(\frac{1}{z} - \frac{1}{v}\right), \tag{5.14}$$

onde $z = \text{sinal}(\hat{\phi})[2\{\hat{\phi}y - K(\hat{\phi})\}]^{1/2}$ e $v = \hat{\phi}K''(\hat{\phi})^{1/2}$, sendo $\hat{\phi}$ obtido de $K'(\hat{\phi}) = y$. A equação (5.14) é usada rotineiramente para aproximar inúmeras funções de distribuição de variáveis aleatórias baseando-se nas suas funções geratrizes de cumulantes.

O uso direto das equações (5.10) e (5.11) para computar probabilidades requeridas na inferência através da verossimilhança envolve integração numérica. Entretanto, aproximações bem mais simples para calcular probabilidades do tipo $P(\hat{\theta} \leq \theta;y)$ são baseadas na aproximação (5.14) redefinindo as quantidades z e v. No caso de θ ser um escalar, Barndorff-Nielsen (1990) e Fraser (1990) integraram a equação (5.10) e deduziram uma fórmula geral análoga à equação (5.14) para calcular a função de distribuição de $\hat{\theta}$ dado

$Y = y$, ou seja, $F_{\hat{\theta}}(\theta; y) = P(\hat{\theta} \leq \theta; y)$, que pode ser expressa como

$$F_{\hat{\theta}}(\theta; y) = \left\{ \Phi(r) + \phi(r) \left(\frac{1}{r} - \frac{1}{u} \right) \right\} \{1 + O(n^{3/2})\} . \tag{5.15}$$

As quantidades em (5.15) são definidas por

$$\begin{aligned} r &= \text{sinal}(\hat{\theta} - \theta)[2\{\ell(\hat{\theta}; y) - \ell(\theta; y)\}]^{1/2}, \\ u &= \left\{ \left. \frac{\partial \ell(\theta; y)}{\partial y} \right|_{\hat{\theta}} - \frac{\partial \ell(\theta; y)}{\partial y} \right\} k(\hat{\theta})^{-1} J(\hat{\theta})^{1/2} \end{aligned} \tag{5.16}$$

com $k(\theta) = \frac{\partial^2 \ell(\theta; y)}{\partial \theta \partial y}$.

Para modelos exponenciais verifica-se de imediato que a quantidade u é igual à estatística de Wald $W = (\hat{\theta} - \theta) J(\hat{\theta})^{1/2}$. Uma forma alternativa para (5.16) segue de

$$F_{\hat{\theta}}(\theta; y) = \Phi(r^*)\{1 + O(n^{-3/2})\}, \tag{5.17}$$

onde $r^* = r + r^{-1} \log(u/r)$. A versão (5.17) pode, algumas vezes, ser mais precisa do que a equação (5.15) embora a diferença seja mínima.

Exemplo 5.7 *Considere a distribuição gama, cuja função densidade é definida por* $\alpha^p y^{p-1} e^{-\alpha y} / \Gamma(p)$ *supondo que o parâmetro de forma* p *é conhecido. O interesse aqui reside no parâmetro* $\theta = \log \alpha$, *que representa um parâmetro de locação. A função de distribuição aproximada de* $\hat{\theta}$ *segue de (5.15) ou (5.17), com as quantidades* r *e* u *obtidas de (5.16) como* $r = [2p\{e^{\hat{\theta}-\theta} - (\hat{\theta} - \theta) - 1\}]^{1/2}$ *e* $u = p^{-1/2}(e^{\hat{\theta}-\theta} - 1)$.

Uma das maiores aplicações práticas das equações (5.15) e (5.17) reside no cálculo de probabilidades associadas à própria função de distribuição da variável aleatória Y proposta para os dados. Essas probabilidades são calculadas através da aproximação $F_Y(y; \theta) = P(Y \leq y; \theta) \doteq P(\hat{\theta} \leq \theta; y)$ com $P(\hat{\theta} \leq \theta; y)$ obtido das equações (5.15) ou (5.17) fazendo $n = 1$. Assim, as probabilidades associadas a variável aleatória Y decorrem daquelas probabilidades relativas a EMV $\hat{\theta}$. A aproximação (5.15) fornece bons resultados na prática, conforme ilustram os exemplos a seguir comparando as aproximações $\Phi(r)$, (5.15) e (5.17) com os valores exatos provenientes da função de distribuição de Y.

Exemplo 5.8 *Suponha uma única observação da distribuição de Cauchy cuja função densidade é* $f(y;\theta) = \pi^{-1}\{1 + (y-\theta)^2\}^{-1}$. *A função de distribuição acumulada exata para* $\theta = 0$ *é* $F(y;0) = 0,5 + \pi^{-1}\text{arctg}\, y$. *Neste caso, com* $\theta = 0$, *obtêm-se de (5.16) as quantidades* $r = \text{sinal}(\hat{\theta})\{2\log(1+\hat{\theta}^2)\}^{1/2}$ *e* $u = \sqrt{2\hat{\theta}}(1+\hat{\theta}^2)^{-1}$, *onde* $\hat{\theta}$ *é calculado iterativamente como descrito no exemplo 1.4. A Tabela 5.1 apresenta as aproximações* $\Phi(r)$, *(5.15) e (5.17) para calcular a função de distribuição* $F(y;0)$. *Com base nesta tabela conclui-se que a equação (5.15) fornece bons resultados para* $F(y;0)$, *enquanto a aproximação* $\Phi(r)$ *não se aplica à distribuição de Cauchy.*

Tabela 5.1: *Probabilidades exatas e aproximadas (expressas em percentagens) para a função de distribuição de Cauchy com* $\theta = 0$

y	-100	-30	-5	-1
exata	0,32	1,06	6,28	25,00
$\Phi(r)$	0;0001	0,01	0,53	11,95
(5.15)	0,28	0,94	5,58	23,22
(5.17)	0,15	0,61	4,69	22,84

Exemplo 5.9 *Considere a distribuição exponencial com média* μ *cuja função de distribuição acumulada é* $P(Y \leq y;\mu) = F_Y(y;\mu) = 1 - e^{-\lambda\mu}$. *A Tabela 5.2 compara os valores exatos de* $P(Y \geq y;\mu)$ *com* $\overline{\Phi}(r) = 1 - \Phi(r)$ *e com aquelas aproximações* $P(\hat{\theta} \geq \theta;y) = 1 - F_{\hat{\theta}}(\theta;y)$ *originárias de (5.15) e (5.17), supondo* $\mu = 1$ *e* $n = 1$. *Observa-se que estas equações fornecem melhores resultados do que a aproximação* $\overline{\Phi}(r)$.

Tabela 5.2: *Valores exatos e aproximados de* $P(Y \geq y;\mu)$ *para a distribuição exponencial com* $\mu = 1$

y	exato	$\overline{\Phi}(r)$	(5.15)	(5.17)
0,5	0,6065	0,7329	0,6070	0,6043
1,0	0,3679	0,5000	0,3695	0,3670
3,0	0,0498	0,0897	0,0505	0,0500
5,0	0,00674	0,0144	0,00689	0,00681
7,0	0,000912	0,00220	0,000938	0,000926
9,0	0,000123	0,000329	0,000127	0,000126

Exemplo 5.10 *Considere dois modelos da família exponencial definidos pelas funções densidade seguintes:*

Modelo	*Função densidade*
log *gama com parâmetro de forma* θ	$\Gamma(\theta)^{-1}\exp(\theta y - e^{y})$
log *gama com parâmetro de locação* θ	$\exp\{(y-\theta) - e^{y-\theta}\}$

Na Tabela 5.3 (Fraser, 1990) as aproximações $\Phi(r)$ *e (5.15) são comparadas com o valor exato de* $P(Y \leq y;\theta)$*, onde valores com* $*$ *se referem às probabilidades complementares* $P(Y \geq y;\theta)$*. Os números desta tabela evidenciam a boa adequação da aproximação (5.15) e sua superioridade em relação a função* $\Phi(r)$*.*

Tabela 5.3: *Probabilidades* $P(Y \leq y;\theta)$ *(expressas em percentagens) para dois modelos log-gama sendo os complementos* $P(Y \geq y;\theta)$ *marcados com* $*$

Modelo log-gama com parâmetro de forma $\theta = 3$

y	$-0,577$	$0,423$	$1,26^{*}$	$1,71^{*}$	$2,14^{*}$
exato	1,95	19,78	31,87	8,79	0,92
$\Phi(r)$	2,73	23,11	28,73	7,62	0,77
(5.15)	1,91	19,61	31,98	8,82	0,93

Modelo log-gama com parâmetro de locação $\theta = 0$

y	-7	-3	-1	1^{*}	2^{*}
exato	0,08	4,86	30,78	6,60	0,06
$\Phi(r)$	0,03	2,14	19,55	11,53	0,15
(5.15)	0,10	5,01	31,04	6,63	0,06

Exemplo 5.11 *Considere dois modelos não pertencentes à família exponencial definidos pelas funções densidade seguintes:*

Modelo	*Função densidade*
gama	$\Gamma(p)^{-1}(y-\theta)^{p-1}e^{-(y-\theta)}$
logístico	$e^{y-\theta}(1+e^{y-\theta})^{-2}$

Os dois modelos são de locação da forma $f(y-\theta)$*. Na Tabela 5.4 (Fraser, 1990) comparam-se as aproximações* $\Phi(r)$ *e (5.15) com os valores exatos de* $P(Y \leq y;\theta)$*. Novamente, a aproximação (5.15) é bastante adequada para calcular a função de distribuição*

de Y e representa um aperfeiçoamento sobre a aproximação $\Phi(r)$, principalmente nas caudas de sua distribuição.

Tabela 5.4: *Probabilidades $P(Y \leq y;\theta)$ (expressas em percentagens) para dois modelos de locação sendo os complementos $P(Y \geq y;\theta)$ marcados com $*$*

Modelo gama com $\theta = 0$ e $p = 3$

y	1	3*	5*	7*	10*
exato	8,03	42,32	12,47	2,96	0,28
$\Phi(r)$	18,97	26,93	6,33	1,28	0,10
(5.15)	7,30	43,28	12,83	3,06	0,29

Modelo logístico com $\theta = 0$

y	-8	-6	-4	-2	-1
exato	0,03	0,25	1,80	11,92	26,89
$\Phi(r)$	0,01	0,12	1,07	9,39	24,41
(5.15)	0,04	0,27	1,87	12,14	27,13

5.6 Correção de Bartlett

Os testes em grandes amostras apresentados na Seção 4.2 são freqüentemente usados na Estatística, pois os testes exatos nem sempre existem. Esses testes são denominados "assintóticos de primeira ordem", isto é, são baseados em valores críticos obtidos de uma distribuição limite conhecida. Um problema natural que surge é verificar se a aproximação de primeira ordem é adequada para a distribuição nula da estatística de teste em consideração. Os testes em grandes amostras, cujas distribuições de referência são qui-quadrado, mais conhecidos são: razão de verossimilhança (w), escore (S_R) e Wald (W). Como foi demonstrado na Seção 4.3, as estatísticas destes três testes são equivalentes em grandes amostras e, em problemas regulares, convergem segundo a hipótese nula H para a distribuição χ^2_q, onde q é o número de restrições impostas por H. Entretanto, em pequenas amostras, a aproximação de primeira ordem pode não ser satisfatória, conduzindo a taxas de rejeição bastante distorcidas. A primeira idéia para melhorar os testes estatísticos foi proposta por Bartlett (1937). Ele considerou apenas a razão de verossimilhança, computando o seu valor esperado segundo H até ordem n^{-1}, onde n é o tamanho da amostra.

Considere um modelo paramétrico $f(y;\theta)$, onde $\theta(\psi^T,\lambda^T)^T$, $\dim(\psi)=q$ e $\dim(\lambda)=p-q$. Deseja-se testar a hipótese nula composta $H:\psi=\psi^{(0)}$ versus $A:\psi\neq\psi^{(0)}$, sendo λ um vetor de parâmetros de perturbação. Seja w a razão de verossimilhança obtida de (4.2). Bartlett propôs calcular o valor esperado de w segundo H até ordem n^{-1} como $E(w)=q+b+O(n^{-2})$, onde $b=b(\psi^{(0)},\lambda)$ é uma constante de ordem $O(n^{-1})$, que pode ser estimada segundo a hipótese nula H. Pode-se verificar, facilmente, que a razão de verossimilhança modificada $w^*=w/(1+b/q)$ tem valor esperado q, exceto por termos de ordem $o(n^{-1})$. O fator de correção $c=1+b/q$ tornou-se conhecido como *correção de Bartlett*, sendo designado para definir uma *razão de verossimilhança aperfeiçoada* que tem distribuição, segundo a hipótese nula, mais próxima da distribuição χ^2_q de referência do que a razão de verossimilhança w usual.

Em problemas regulares, para testar uma hipótese nula composta qualquer, Lawley (1956) deduziu uma fórmula geral para b em termos de *cumulantes da log-verossimilhança*, que são simplesmente valores esperados de produtos de derivadas da log-verossimilhança. Além disso, através de uma demonstração extremamente complicada, Lawley concluiu que os momentos de w^* concordam com aqueles correspondentes da distribuição χ^2_q exceto por termos de ordem n^{-2}. Este resultado é muito importante, pois mostra que a simples correção do primeiro momento de w possibilita obter um teste aperfeiçoado baseado em w^*, cujos momentos (segundo H) concordam, até termos de ordem n^{-1}, com aqueles correspondentes da distribuição qui-quadrado de referência.

Hayakawa (1977) apresenta a expansão da função densidade de w até $O(n^{-1})$ supondo a hipótese nula $H:\psi=\psi^{(0)}$ verdadeira que, após simplificações conduzidas por Cordeiro (1987) e Chesher e Smith (1995), pode ser expressa como

$$f_w(x)=f_q(x)\left\{1+\frac{b}{2}\left(\frac{x}{q}-1\right)\right\}, \tag{5.18}$$

onde, de agora por diante, $f_q(x)$ representa a função densidade da variável aleatória χ^2_q. Note-se que $f_w(x)$ só depende da dimensão de ψ, da função densidade $f_q(x)$ da distribuição χ^2_q de referência e do termo de ordem n^{-1} em $E(w)$. De (5.18) é fácil mostrar que a função densidade de $w^*=w/(1+b/q)$ ou $w(1-b/q)$, segundo H e até termos de ordem $O(n^{-1})$, é $f_{w^*}(x)=f_q(x)$, o que comprova que a razão de verossimilhança modificada

pela correção de Bartlett tem distribuição idêntica à distribuição χ^2_q, exceto por termos de ordem $O(n^{-2})$, como primeiro estabelecido por Lawley. Observa-se que (5.18) é uma expansão do tipo (3.1) pois a constante b é de ordem $O(n^{-1})$. Assim, enquanto $P(w \leq x) = P(\chi^2_q \leq x) + O(n^{-1})$ tem-se o melhoramento $P(w^* \leq x) = P(\chi^2_q \leq x) + O(n^{-2})$. O erro da aproximação χ^2_q para a distribuição de w é de ordem n^{-1}, enquanto o erro desta aproximação para a distribuição de w^* é reduzido para ordem n^{-2}.

Pode-se escrever w na equação (4.23) do teste de $H : \psi = \psi^{(0)}$ versus $A : \psi \neq \psi^{(0)}$ como

$$w = 2\{\ell(\hat{\psi}, \hat{\lambda}) - \ell(\psi^{(0)}, \tilde{\lambda})\} - 2\{\ell(\psi^{(0)}, \tilde{\lambda}) - \ell(\psi^{(0)}, \lambda)\},$$

onde $\ell(\psi^{(0)}, \lambda)$ é a log-verossimilhança avaliada no parâmetro verdadeiro e $\tilde{\lambda}$ como antes é a EMV de λ restrita a $\psi = \psi^{(0)}$. Lawley (1956) demonstrou que

$$2E\{\ell(\hat{\psi}, \hat{\lambda}) - \ell(\psi^{(0)}, \lambda)\} = p + \epsilon_p, \tag{5.19}$$

onde ϵ_p é um termo de ordem n^{-1} dado por

$$\epsilon_p = \sum(\ell_{rstu} - \ell_{rstuvw}), \tag{5.20}$$

sendo que $\sum$ é o somatório sobre todas as componentes do vetor θ, isto é, os índices r, s, t, u, v e w variam sobre os p parâmetros, e os $\ell's$ têm expressões

$$\begin{aligned} \ell_{rstu} = &\ \kappa^{rs}\kappa^{tu}\left\{\kappa_{rstu}/4 - \kappa^{(u)}_{rst} + \kappa^{(tu)}_{rs}\right\}, \\ \ell_{rstuvw} = &\ \kappa^{rs}\kappa^{tu}\kappa^{vw}\{\kappa_{rtv}(\kappa_{suw}/6 - \kappa^{(u)}_{sw}) \\ &+\kappa_{rtu}(\kappa_{svw}/4 - \kappa^{(v)}_{sw}) + \kappa^{(v)}_{rt}\kappa^{(u)}_{sw} + \kappa^{(u)}_{rt}\kappa^{(v)}_{sw}\}\ , \end{aligned} \tag{5.21}$$

onde os cumulantes $\kappa's$ são definidos na Seção 5.2. A matriz de informação total de Fisher para θ tem elementos $\kappa_{r,s} = -\kappa_{rs}$, sendo $\kappa^{r,s} = -\kappa^{rs}$ os correspondents elementos de sua inversa. Os $\ell's$ das equações em (5.21) são, em geral, de ordem n^{-1}. O valor esperado de $2\{\ell(\psi^{(0)}, \tilde{\lambda}) - \ell(\psi^{(0)}, \lambda)\}$ segue expressão análoga àquela de $2\{\ell(\hat{\psi}, \hat{\lambda}) - \ell(\psi^{(0)}, \lambda)\}$, ou seja, $2E\{\ell(\psi^{(0)}, \tilde{\lambda}) - \ell(\psi^{(0)}, \lambda)\} = p - q + \epsilon_{p-q} + O(n^{-2})$, com ϵ_{p-q} deduzido da equação (5.20) observando, agora, que o somatório Σ daquela fórmula se estende apenas sobre as componentes em λ, isto é, sobre os $p - q$ parâmetros de perturbação, uma vez que ψ está fixo em $\psi^{(0)}$.

Então, segundo H, o valor esperado da razão de verossimilhança é $E(w) = q + \epsilon_p - \epsilon_{p-q} + O(n^{-2})$ e, portanto, pode-se melhorar a aproximação da estatística de teste pela distribuição χ^2_q trabalhando com $w^* = w/c$, ao invés de w, onde a correção de Bartlett é obtida de

$$c = 1 + \frac{\epsilon_p - \epsilon_{p-q}}{q} . \tag{5.22}$$

A estatística corrigida w^* tem distribuição χ^2_q até $O(n^{-1})$ sob H. Em outras palavras, o teste aperfeiçoado compara w^* com a distribuição χ^2_q de referência. A dificuldade do aperfeiçoamento reside no cálculo de ε_p e ε_{p-q} a partir das equações (5.20) e (5.21). No caso da correção de Bartlett depender de parâmetros desconhecidos, eles devem ser substituídos pelas suas estimativas de máxima verossimilhança segundo H, mas isto não afeta a ordem da aproximação resultante. O inconveniente no uso da fórmula de Lawley (5.22) na prática é o cálculo do grande número de produtos de cumulantes em testes envolvendo três ou mais parâmetros. Entretanto, para vários modelos estatísticos, os cumulantes da log-verossimilhança são invariantes segundo permutação de parâmetros, conforme descrito por Cordeiro (1983) no contexto dos modelos lineares generalizados.

No caso uniparamétrico relativo ao teste de $H : \theta = \theta^{(0)}$ versus $A : \theta \neq \theta^{(0)}$, a correção de Bartlett para a razão de verossimilhança $w = 2\{\ell(\hat{\theta}) - \ell(\theta^{(0)})\}$ é deduzida de (5.20) – (5.21), fazendo todos os índices iguais a θ, implicando

$$\epsilon_1 = \kappa^{\theta\theta^2}\{\kappa_{\theta\theta\theta}/4 - \kappa^{(\theta)}_{\theta\theta\theta} + \kappa^{(\theta\theta)}_{\theta\theta}\} - \kappa^{\theta\theta^3}\{\kappa_{\theta\theta\theta}(5\kappa_{\theta\theta\theta}/12 - 2\kappa^{(\theta)}_{\theta\theta}) + 2\kappa^{(\theta)^2}_{\theta\theta}\} . \tag{5.23}$$

A razão de verossimilhança modificada pela correção $1 + \epsilon_1$, i.e., $w^* = w/(1 + \epsilon_1)$, tem distribuição nula aproximada pela distribuição χ^2_1 com erro $O(n^{-2})$.

Uma metodologia para calcular as correções de Bartlett em modelos estatísticos consiste em: (i) inverter a matriz de informação segundo H e A; (ii) calcular os cumulantes $\kappa's$ que aparecem em (5.21) para todas as combinações de parâmetros; (iii) substituir os $\kappa's$ em (5.21) e desenvolver as somas em (5.20) sobre todos os parâmetros em θ e sobre aqueles parâmetros de perturbação em λ; (iv) manipular os termos em ϵ_p e ϵ_{p-q} com o intuito de encontrar expressões algébricas simples. A reparametrização, quando possível, visando ortogonalizar os vetores de parâmetros λ e ψ (Seção 4.3) implica grandes simplificações no cálculo das correções de Bartlett.

Exemplo 5.12 *Considere n observações iid da distribuição* $N(\mu, \sigma^2)$. *O interesse reside em calcular as correções de Bartlett para os testes de* $H_1 : \mu = \mu^{(0)}$ *versus* $A_1 : \mu \neq \mu^{(0)}$ (σ^2 *desconhecido) e* $H_2 : \sigma^2 = \sigma^{(0)^2}$ *versus* $A_2 : \sigma^2 \neq \sigma^{(0)^2}$ (μ *desconhecido). As estatísticas da razão de verossimilhança para estes testes são obtidas da log-verossimilhança* $\ell(\mu, \sigma^2)$ *sendo dadas por*

$$w_1 = 2\{\ell(\hat{\mu}, \hat{\sigma}^2) - \ell(\mu^{(0)}, \tilde{\sigma}^2)\} = n \log \left\{ \frac{\Sigma(y_i - \mu^{(0)})^2}{\Sigma(y_i - \overline{y})^2} \right\}$$

e

$$w_2 = 2\{\ell(\hat{\mu}, \hat{\sigma}^2) - \ell(\tilde{\mu}, \sigma^{(0)^2})\} = n \left[\log \left(\frac{\sigma^{(0)^2}}{\hat{\sigma}^2} \right) + \frac{\hat{\sigma}^2 - \sigma^{(0)^2}}{\sigma^{(0)^2}} \right] ,$$

respectivamente, onde $\hat{\mu} = \tilde{\mu} = \overline{y}$, $\hat{\sigma}^2 = \Sigma(y_i - \overline{y})^2/n$ *e* $\tilde{\sigma}^2 = \Sigma(y_i - \mu^{(0)})^2/n$. *Os cumulantes* κ*'s para o cálculo das correções de Bartlett são então deduzidos como no exemplo 5.1. Usando as equações (5.20) e (5.21) pode-se obter* $E(w_1)$ *e* $E(w_2)$ *até* $O(n^{-1})$ *considerando as somas sobre todas as componentes de* $\theta = (\mu, \sigma^2)^T$ *e fazendo todos os índices iguais ao parâmetro* σ^2 *e ao parâmetro* μ, *respectivamente. Assim,*

$$E(w_1) = 1 + \sum_{\mu,\sigma^2} (\ell_{rstu} - \ell_{rstuvw}) - (\ell_{\sigma^2\sigma^2\sigma^2\sigma^2} - \ell_{\sigma^2\sigma^2\sigma^2\sigma^2\sigma^2\sigma^2})$$

e

$$E(w_2) = 1 + \sum_{\mu,\sigma^2} (\ell_{rstu} - \ell_{rstuvw}) - (\ell_{\mu\mu\mu\mu} - \ell_{\mu\mu\mu\mu\mu\mu}) .$$

Computando-se os ℓ*'s e após alguma álgebra obtêm-se*

$$E(w_1) = 1 + 3/(2n) \quad e \quad E(w_2) = 1 + 11/(6n),$$

de onde seguem as estatísticas modificadas $w_1^* = w_1/(1 + 3/2n)$ *e* $w_2^* = w_2/(1 + 11/6n)$ *para melhorar os testes de* H_1 *e* H_2, *respectivamente. Aqui, as correções de Bartlett não dependem de parâmetros desconhecidos. Elas podem ser obtidas por primeiros princípios dos resultados* $n\hat{\sigma}^2/\sigma^2 \sim \chi^2_n$ *e* $n\tilde{\sigma}^2/\sigma^2 \sim \chi^2_{n-1}$ *aproximando* $E(\log \chi^2_n)$ *por* $\log n - n^{-1}$.

Exemplo 5.13 *Considere n observações iid de uma distribuição exponencial com média* μ. *A log-verossimilhança para* μ *é* $\ell(\mu) = -n \log \mu - n\overline{y}/\mu$, *onde* $\overline{y}$ *é média das observações. A razão de verossimilhança para testar* $H : \mu = \mu^{(0)}$ *versus* $A : \mu \neq \mu^{(0)}$ *é dada*

por $w = 2n\{\overline{y}\log(\overline{y}/\mu^{(0)}) - (\overline{y} - \mu^{(0)})\}$. *Os cumulantes seguem de* $\kappa_{\mu,\mu} = n/\mu^2$, $\kappa_{\mu,\mu,\mu} = -\kappa_{\mu,\mu\mu} = 2n/\mu^3$, $\kappa_{\mu\mu\mu} = 4n/\mu^3$, $\kappa_{\mu\mu\mu\mu} = -30n/\mu^4$, $\kappa_{\mu,\mu\mu\mu} = 18n/\mu^4$, *etc. Substituindo estes cumulantes em (5.23) obtém-se a correção de Bartlett* $c = 1 + \epsilon_1$ *como* $c = 1 + 1/(6n\mu^{(0)})$.

As pesquisas em correções de Bartlett tiveram um grande impulso a partir de 1982 e, nos dias atuais, constituem uma área consolidada de grande interesse da teoria assintótica. Estas pesquisas seguem quatro direções principais: a primeira corresponde ao desenvolvimento de fórmulas algébricas simples para as correções em modelos especiais; a segunda pesquisa métodos alternativos gerais de cálculo das correções de Bartlett; a terceira restringe-se a aplicações numéricas e a estudos de simulação; finalmente, a quarta visa a interpretar as correções à luz da geometria diferencial e a relacioná-las com tópicos de interesse recente, como ortogonalidade de parâmetros, verossimilhanças não-canônicas, etc.

Cordeiro (1983, 1987) e Cordeiro e Paula (1989) desenvolveram fórmulas gerais para as correções de Bartlett em notação matricial nos modelos lineares generalizados e nos modelos não-lineares da família exponencial, respectivamente. Barndorff-Nielsen e Cox (1984a) apresentaram um método indireto de cálculo das correções de Bartlett em modelos paramétricos gerais, a partir de uma simples relação entre a correção e as constantes normalizadoras da distribuição da estimativa de máxima verossimilhança condicional a uma estatística ancilar, exata ou aproximada. Também, Barndorff-Nielsen e Cox (1984b) investigaram a distribuição de w com relação a vários tipos de censura e regras de parada nos processos Browniano e de Poisson. Porteous (1985) obteve correções para modelos de seleção de covariáveis quando a razão de verossimilhança tem forma fechada. Correções de Bartlett para testes em modelos multivariados com matrizes de covariância estruturais foram desenvolvidos por Møller (1986). McCullagh e Cox (1986) interpretaram a correção de Bartlett em termos de combinações invariantes de cumulantes das duas primeiras derivadas da log-verossimilhança. Barndorff-Nielsen e Blaesild (1986) propuseram um algoritmo para calcular as correções em situações onde várias hipóteses alternativas são lineares na mesma parametrização. Uma forma para $E(w)$ invariante em relação a permutação de parâmetros foi desenvolvida para modelos exponenciais por Ross (1987).

Attfield (1991) e Cordeiro (1993) mostraram como corrigir os testes da razão de verossimilhança em modelos heterocedásticos. Cordeiro, Paula e Botter (1994) obtiveram correções de Bartlett para a classe dos modelos de dispersão proposta por Jørgensen (1987), generalizando os resultados de Cordeiro (1983, 1987) e Cordeiro e Paula (1989). Finalmente, Cordeiro et al. (1995) apresentaram fórmulas gerais simples para as correções de Bartlett em modelos exponenciais uniparamétricos.

5.7 Estatísticas Aperfeiçoadas tendo distribuição χ^2

Como foi apresentado nas Seções 4.2 e 4.3, os testes escore e de Wald são assintoticamente equivalentes aos testes baseados na razão de verossimilhança. Cordeiro e Ferrari (1991) demonstraram que, sob condições gerais de regularidade como aquelas descritas na Seção 4.2, qualquer estatística S cuja distribuição assintótica é qui-quadrado pode ser aperfeiçoada por um fator de correção multiplicativo expresso como um *polinômio* de grau k, de modo que os momentos da estatística modificada sejam idênticos aos correspondentes da distribuição qui-quadrado de referência, exceto por termos de ordem n^{-2}. A estatística corrigida tem a forma $S^* = S(1-\sum_{i=1}^{k} c_i S^{i-1})$, onde os $c_i's$ de ordem n^{-1} são determinados de tal maneira que a distribuição de S^* sob a hipótese nula seja qui-quadrado (até esta ordem). O fator multiplicativo acima é denominado *correção tipo-Bartlett*, sendo uma extensão da clássica correção de Bartlett correspondente ao caso de $k = 1$. Apresenta-se agora a demonstração deste resultado.

Seja S uma estatística arbitrária com a finalidade de testar uma hipótese nula composta cuja distribuição assintótica, supondo esta hipótese verdadeira, é qui-quadrado com q graus de liberdade, ou seja, sua função de distribuição $F_S(x)$ satisfaz $\lim_{n\to\infty} F_S(x) = F_q(x)$, onde $F_q(x)$ representa a função de distribuição da variável χ^2_q. Sob certas condições de regularidade, Chandra (1985) demonstrou que $F_S(x)$ pode ser expressa até $O(n^{-1})$ como uma combinação linear finita de funções de distribuição qui-quadrado com graus de liberdade $q, q+2, \ldots, q+2k$. Assim, a função de distribuição de S, na qual termos de ordem inferior a n^{-1} são omitidos, pode ser escrita como

$$F_S(x) = F_q(x) + \sum_{i=0}^{k} a_i F_{q+2i}(x), \tag{5.24}$$

onde os $a_i's$ são quantidades de ordem n^{-1}. Na realidade, elas são funções de parâmetros desconhecidos. Para que a função $F_S(x)$ em (5.24) seja uma função de distribuição até ordem $O(n^{-1})$ é necessário que a condição $\sum_{i=0}^{k} a_i = 0$ seja satisfeita. As estatísticas escore e de Wald apresentam expansões do tipo (5.24) para suas funções de distribuição com $k = 3$, enquanto $k = 1$ para a razão de verossimilhança.

Sejam as relações de recorrência

$$F_{q+2}(x) = F_q(x) - \frac{2x}{q} f_q(x)$$

e

$$f_{q+2}(x) = \frac{x}{q} f_q(x),$$

onde $f_q(x) = \frac{dF_q(x)}{dx}$ é a função densidade da variável χ^2_q. Usando estas relações, a equação (5.24) pode ser dada por

$$F_S(x) = F_q(x) - f_q(x) \sum_{i=1}^{k} C_i x^i,$$

onde $C_i = 2\mu_i'^{-1} \sum_{\ell=i}^{k} a_\ell$ para $i = 1, \ldots, k$ e

$$\mu_i' = E\{(\chi^2_q)^i\} = 2^i \Gamma\left(\frac{q}{2} + i\right) / \Gamma\left(\frac{q}{2}\right) .$$

A forma funcional anterior, envolvendo um polinômio de grau k, sugere a estatística modificada

$$S^* = S\left(1 - \sum_{i=1}^{k} c_i S^{i-1}\right) . \tag{5.25}$$

Os $c_i's$ são determinados em (5.25) de maneira a satisfazer $F_{S^*}(x) = F_q(x)$ até $O(n^{-1})$, i.e., de modo que S^* tenha sob a hipótese nula distribuição χ^2_q até esta ordem. O teorema de Cox e Reid (Seção 3.7) aplicado à expressão (5.25) produz a função de distribuição de S^* até ordem n^{-1} como

$$F_{S^*}(x) = F_S(x) + f_S(x) \sum_{i=1}^{k} c_i x^i,$$

onde $f_S(x) = \frac{dF_S(x)}{dx}$. Uma vez que S tem distribuição χ^2_q até $O(n^{-1})$, e que os $c_i's$ são

$O(n^{-1})$, obtém-se até esta ordem

$$F_{S^*}(x) = F_S(x) + f_q(x)\sum_{i=1}^{k} c_i x^i \ . \tag{5.26}$$

Substituindo na equação (5.26) a expansão de $F_S(x)$ dada anteriormente tem-se

$$F_{S^*}(x) = F_q(x) + f_q(x)\sum_{i=1}^{n}(c_i - C_i)x^i \ .$$

A igualdade $F_{S^*}(x) = F_q(x)$ até ordem n^{-1} é satisfeita se, e somente se, $c_i = C_i$ para $i = 1, \ldots, k$. Consequentemente, a estatística aperfeiçoada

$$S^* = S\left\{1 - 2\sum_{i=1}^{k}\left(\sum_{\ell=i}^{k} a_\ell\right)\mu_i'^{-1}S^{i-1}\right\} \tag{5.27}$$

tem distribuição χ^2_q até ordem n^{-1} sob a hipótese nula. O termo entre chaves na fórmula (5.27) é denominado *correção tipo-Bartlett* e objetiva melhorar a aproximação da distribuição da estatística S^* pela distribuição χ^2_q. O melhoramento é no sentido de que $P(S^* \leq x) = F_q(x) + O(n^{-2})$ enquanto $P(S \leq x) = F_q(x) + O(n^{-1})$, ou seja, baseando-se o teste em S^*, o erro da aproximação qui-quadrado é reduzido de $O(n^{-1})$ para $O(n^{-2})$. A correção tipo-Bartlett quando $k > 1$ não é uma correção de Bartlett genuína, pois envolve a própria estatística não-modificada. Claramente, no caso da razão de verossimilhança, quando $k = 1$, a correção em (5.27) se torna igual a um escalar que é a própria correção de Bartlett.

Os coeficientes $a_1, \ldots, a_k$ necessários para se obter S^* podem ser expressos como funções dos termos de ordem $O(n^{-1})$ dos k primeiros momentos da estatística não-modificada S (vide Cordeiro e Ferrari, 1998). Estes coeficientes são calculados para cada tipo de estatística (razão de verossimilhança, escore, Wald, Wald modificada, etc.) através de fórmulas especiais como funções dos cumulantes conjuntos $\kappa's$ (vide Seção 5.2). Fórmulas matriciais para os $a_i's$ relativas aos testes escore são dadas, em generalidade, por Ferrari e Cordeiro (1994).

5.8 Testes Escore Melhorados

Os testes escore, também conhecidos como testes do multiplicador de Lagrange, são bastante usados em Estatística e Econometria como uma alternativa para os testes da razão de verossimilhança, principalmente quando a estimação segundo a hipótese alternativa é mais trabalhosa do que segundo a hipótese nula. Neste caso, os testes escore são mais simples pois requerem somente estimação segundo a hipótese nula. Em tabelas de contingência para análise de dados sob a forma de contagens, os testes usuais conhecidos como χ^2 de Pearson são testes escore. As aplicações dos testes escore aparecem em modelos lineares generalizados (Pregibon, 1982), em modelos de séries temporais (Hosking, 1980, 1981 e Poskitt e Tremayne, 1981, 1982), em modelos de sobrevivência (Lawless, 1982) e em inúmeros modelos econométricos (Breusch e Pagan, 1980 e Engle, 1984).

Retorna-se aqui ao problema descrito na Seção 4.3 de testar a hipótese nula composta $H : \psi = \psi^{(0)}$ versus a hipótese alternativa composta $A : \psi \neq \psi^{(0)}$, onde $\theta = (\psi^T, \lambda^T)^T$, $\dim(\psi) = q$ e $\dim(\lambda) = p - q$. A função escore total $U(\theta) = (U_\psi(\psi, \lambda)^T, U_\lambda(\psi, \lambda)^T)^T$ para θ é particionada conforme θ. A matriz de informação $K = K(\theta)$ para θ e sua inversa, particionadas como θ, são

$$K(\theta) = \begin{pmatrix} K_{\psi\psi} & K_{\psi\lambda} \\ K_{\lambda\psi} & K_{\lambda\lambda} \end{pmatrix} \quad \text{e} \quad K(\theta)^{-1} = \begin{pmatrix} K^{\psi\psi} & K^{\psi\lambda} \\ K^{\lambda\psi} & K^{\lambda\lambda} \end{pmatrix},$$

onde todas as submatrizes acima são, em geral, funções de ψ e λ. Sejam $\hat{\theta} = (\hat{\psi}^T, \hat{\lambda}^T)$ e $\tilde{\theta} = (\psi^{(0)T}, \tilde{\lambda}^T)^T$ as EMV irrestrita e restrita de θ, respectivamente. As funções avaliadas em $\tilde{\theta}$ são, como antes, denotadas com um til. A estatística escore S_R para testar $H : \psi = \psi^{(0)}$ versus $A : \psi \neq \psi^{(0)}$ pode ser expressa como $S_R = \tilde{U}_\psi^T \tilde{K}^{\psi\psi} \tilde{U}_\psi$, onde $\tilde{U}_\psi = U_\psi(\tilde{\psi}, \tilde{\lambda})$. Como foi estabelecido na Seção 4.3, satisfeitas certas condições de regularidade como aquelas da Seção 4.1.3, a distribuição de S_R converge em grandes amostras para a distribuição χ^2_q sob a hipótese nula.

A expansão assintótica da função de distribuição de S_R segue a expansão (5.24) com $k = 3$ (Harris, 1985). Para apresentar os seus coeficientes a_0, a_1, a_2 e a_3, necessita-se definir as seguintes matrizes particionadas conforme θ:

$$A = \begin{pmatrix} 0 & 0 \\ 0 & K_{\lambda\lambda}^{-1} \end{pmatrix} \quad \text{e} \quad M = K^{-1} - A,$$

onde $K_{\lambda\lambda}^{-1}$ representa a estrutura de covariância assintótica de $\tilde{\lambda}$. Os elementos típicos

(i,j) de A e M são denotados por a_{ij} e m_{ij}, respectivamente. Harris (1985) demonstrou que

$$a_0 = (A_2 - A_1 - A_3)/24,\ a_1 = (3A_3 - 2A_2 + A_1)/24,$$

$$a_2 = (A_2 - 3A_3)/24 \quad \text{e} \quad a_3 = A_3/24,$$

onde as quantidades A_1, A_2 e A_3 de ordem n^{-1} são dadas como funções dos cumulantes conjuntos $\kappa's$ (Seção 5.2) por

$$\begin{aligned}
A_1 &= 3\Sigma(\kappa_{ijk} + 2\kappa_{i,jk})(\kappa_{rst} + 2\kappa_{rs,t})a_{ij}\ a_{st}\ m_{kr} \\
&\quad -6\Sigma(\kappa_{ijk} + 2\kappa_{i,jk})\kappa_{r,s,t}\ a_{ij}\ a_{kr}\ m_{st} \\
&\quad +6\Sigma(\kappa_{i,jk} - \kappa_{i,j,k})(\kappa_{rst} + 2\kappa_{rs,t})a_{js}\ a_{kt}\ m_{ir} \\
&\quad -6\Sigma(\kappa_{i,j,k,r} + \kappa_{i,j,kr})a_{kr}\ m_{ij}, \\
A_2 &= -3\Sigma\kappa_{i,j,k}\ \kappa_{r,s,t}\ a_{kr}\ m_{ij}\ m_{st} \\
&\quad +6\Sigma(\kappa_{ijk} + 2\kappa_{i,jk})\kappa_{r,s,t}\ a_{ij}\ m_{kr}\ m_{st} \\
&\quad -6\Sigma\kappa_{i,j,k}\ \kappa_{r,s,t}\ a_{kt}\ m_{ir}\ m_{js} \\
&\quad +3\Sigma\kappa_{i,j,k,r}\ m_{ij}\ m_{kr}, \\
A_3 &= 3\Sigma\kappa_{i,j,k}\ \kappa_{r,s,t}\ m_{ij}\ m_{kr}\ m_{st} + 2\Sigma\kappa_{i,j,k}\ \kappa_{r,s,t}\ m_{ir}\ m_{js}\ m_{kt}.
\end{aligned} \tag{5.28}$$

As somas nas equações (5.28) são tomadas em relação a todos os parâmetros $\theta_1, \ldots, \theta_p$ de θ. Observe-se que, como esperado, $\sum_{i-0}^{3} a_i = 0$. As fórmulas (5.28) são extremamente complicadas para serem analisadas num contexto geral. Para modelos especiais, elas podem sofrer redução considerável.

Determinando-se os $A_i's$ para o modelo em consideração, a estatística escore aperfeiçoada tem a representação (5.27), ou seja,

$$S_R^* = S_R\{1 - (c + bS_R + aS_R^2)\}, \tag{5.29}$$

onde

$$\begin{aligned}
a &= \frac{A_3}{12q(q+2)(q+4)}, \\
b &= \frac{A_2 - 2A_3}{12q(q+2)}, \\
c &= \frac{A_1 - A_2 + A_3}{12q}.
\end{aligned} \tag{5.30}$$

A correção tipo-Bartlett em (5.29) para melhorar o teste de $H : \psi = \psi^{(0)}$ tem os coeficientes determinados pelas equações (5.30) e (5.28) como funções de cumulantes conjuntos de derivadas da log-verossimilhança. O teste escore aperfeiçoado de $H : \psi = \psi^{(0)}$ pode ser conduzido comparando a estatística escore modificada S_R^* com a distribuição χ_q^2 de referência, sendo o erro da aproximação qui-quadrado de ordem $O(n^{-2})$. No caso das quantidades A_1, A_2 e A_3 envolverem parâmetros em λ, estes devem ser substituídos pelas suas estimativas em $\tilde{\lambda}$ mas o erro da aproximação χ_q^2 para a distribuição nula de S_R^* continuará sendo de ordem $O(n^{-2})$ (Cordeiro e Ferrari, 1991).

Da expansão da função de distribuição de S_R até $O(n^{-1})$, Harris (1985) deduziu até esta ordem e sob $H : \psi = \psi^{(0)}$, os três primeiros momentos de S_R como

$$\begin{aligned}
\mu_1'(S_R) &= q + \frac{A_1}{12}, \\
\mu_2'(S_R) &= q(q+2) + \frac{A_1(q+2) + 2A_2}{6}, \\
\mu_3'(S_R) &= q(q+2)(q+4) + \frac{A_1(q+2)(q+4) + 4A_2(q+4) + 8A_3}{4}.
\end{aligned} \tag{5.31}$$

As equações (5.31) podem ser usadas para calcular A_1, A_2 e A_3 quando os momentos $\mu_r'(S_R)$ de S_R para $r = 1, 2$ e 3 forem mais facilmente determinados por primeiros princípios.

Suponha agora o caso uniparamétrico de testar $H : \theta = \theta^{(0)}$ versus $A : \theta \neq \theta^{(0)}$, onde a estatística escore tem expressão $S_R = [U(\theta)^2 / E\{U(\theta)^2\}]_{\theta=\theta^{(0)}}$ sendo $U(\theta) = d\ell(\theta)/d\theta$ a função escore total para θ com o quociente em S_R avaliado em $\theta = \theta^{(0)}$. Para melhorar o teste de H demonstra-se (Cordeiro e Ferrari, 1991) que as quantidades A_1, A_2 e A_3 em (5.28) são dadas por $A_1 = 0$, $A_2 = 3\kappa_4/\kappa_2^2$ e $A_3 = 5\kappa_3^2/\kappa_2^3$, onde $\kappa_2 = E\{U(\theta)^2\}$ é a informação total de Fisher para θ e $\kappa_3 = E\{U(\theta)^3\}$ e $\kappa_4 = E\{U(\theta)^4\} - 3\kappa_2^2$ são os terceiro e quarto cumulantes da função escore total, respectivamente. Sejam $\gamma_1 = \kappa_3/\kappa_2^{3/2}$ e $\gamma_2 = \kappa_4/\kappa_2^2$ as medidas usuais de assimetria e curtose da função escore, isto é, os seus terceiro e quarto cumulantes padronizados. A estatística escore aperfeiçoada (5.29) para testar $H_0 : \theta = \theta^{(0)}$ tem a forma simples

$$S_R^* = S_R \left[1 - \frac{1}{36} \left\{ 3(5\gamma_1^2 - 3\gamma_2) + (3\gamma_2 - 10\gamma_1^2) S_R + \gamma_1^2 S_R^2 \right\} \right]. \tag{5.32}$$

O primeiro coeficiente em (5.32), $(5\gamma_1^2 - 3\gamma_2)/12$, é uma medida da não-normalidade ou não-normalidade inversa da função escore, pois se anula para as distribuições normal e normal inversa. O terceiro coeficiente, $\gamma_1^2/36$, corrige a assimetria da função escore e o segundo, $(3\gamma_2 - 10\gamma_1^2)/36$, é uma combinação linear das medidas de assimetria e curtose desta função.

Exemplo 5.14 *Consideram-se aqui três modelos biparamétricos: a distribuição normal* $N(\mu, \phi^{-1})$ *com média* μ *e variância* ϕ^{-1} *e as distribuições normal inversa* $N^-(\mu, \phi)$ *de média* μ *positiva e parâmetro de precisão* ϕ *positivo e gama* $G(\mu, \phi)$ *de média* μ *positiva e índice* ϕ *positivo. As duas últimas distribuições têm as seguintes funções densidades:*

Distribuição	*Função Densidade*
$N^-(\mu, \phi)$	$\left(\frac{\phi}{2\pi y^3}\right)^{1/2} \exp\left\{\frac{-\phi(y-\mu)^2}{2\mu^2 y}\right\}$
$G(\mu, \phi)$	$\left(\frac{\phi}{\mu}\right)^{\phi} y^{\phi-1} e^{-\phi y/\mu}/\Gamma(\phi)$

Para estes três modelos o interesse reside em testar a média $H_1 : \mu = \mu^{(0)}$ *versus* $A_1 : \mu \neq \mu^{(0)}$ *quando o parâmetro de dispersão* ϕ^{-1} *é desconhecido. O cálculo dos cumulantes conjuntos* κ*'s e dos* A_i'*s das equações (5.28) pode ser encontrado em Cordeiro e Ferrari (1991). Apresentam-se, a seguir, as formas das estatísticas escore tradicional* S_R *e aperfeiçoada* S_R^* *nestes três modelos:*

Modelo normal $N(\mu, \phi^{-1})$*:*

$$S_R = \frac{n^2(\overline{y} - \mu^{(0)})^2}{\sum_{i=1}^{n}(y_i - \mu^{(0)})^2} \quad e \quad S_R^* = S_R\left\{1 - \frac{1}{2n}(3 - S_R)\right\};$$

Modelo normal inverso $N^-(\mu, \phi)$*:*

$$S_R = \frac{n^2(\overline{y} - \mu^{(0)})^2}{\mu^{(0)} \sum_{i=1}^{n} \frac{(y_i - \mu^{(0)})^2}{y_i}} \quad e \quad S_R^* = S_R\left[1 - \frac{1}{4n}\left\{6 - \left(2 + \frac{5\mu^{(0)}}{\tilde{\phi}}\right) S_R + \frac{\mu^{(0)}}{\tilde{\phi}} S_R^2\right\}\right],$$

onde $\tilde{\phi} = \dfrac{n\mu^{(0)2}}{\sum_{i=1}^{n}(y_i - \mu^{(0)})^2/y_i}$;

Modelo gama $G(\mu,\phi)$: $S_R = n\tilde{\phi}(\overline{y}-\mu^{(0)})^2/\mu^{(0)^2}$ *e* S_R^* *segue de (5.29) - (5.30) com*

$$A_1 = 6(1-\tilde{\phi}^2\tilde{\psi}'' - 2\tilde{\phi}\tilde{\psi}')/\{n\tilde{\phi}(1-\tilde{\phi}\tilde{\psi}')^2\},$$

$$A_2 = 18(n\tilde{\phi})^{-1} + 9/\{n\tilde{\phi}(1-\tilde{\phi}\tilde{\psi}')\} \quad e \quad A_3 = 20/(n\tilde{\phi}),$$

onde $\tilde{\psi}'$ *e* $\tilde{\psi}''$ *são as derivadas da função digama* ψ *avaliadas na EMV* $\tilde{\phi}$ *restrita que é decorrente da equação*

$$\log\tilde{\phi} - \psi(\tilde{\phi}) = \frac{\overline{y}-\mu^{(0)}}{\mu^{(0)}} + \log\left(\frac{\mu^{(0)}}{\tilde{y}}\right),$$

sendo $\tilde{y}$ *a média geométrica dos dados.*

Exemplo 5.15 *Trabalha-se ainda com os três modelos descritos no exemplo 5.14, onde o interesse agora é testar o parâmetro de precisão* $H_2 : \phi = \phi^{(0)}$ *versus* $A_2 : \phi \neq \phi^{(0)}$ *quando a média* μ *é desconhecida. Apresentam-se a seguir as formas das estatísticas* S_R *e* S_R^* *nestes modelos:*

Modelo normal $N(\mu,\phi^{-1})$:

$$S_R = \frac{1}{2n}\{n - \phi^{(0)}\sum_{i=1}^{n}(y_i-\overline{y})^2\} \quad e$$

$$S_R^* = S_R\left\{1 - \frac{1}{18n}(33 - 34S_R + 4S_R^2)\right\};$$

Modelo normal inverso $N^{-}(\mu,\phi)$:

$$S_R = \frac{1}{2n}\left\{n - \frac{\phi^{(0)}}{\overline{y}^2}\sum_{i=1}^{n}\frac{(y_i-\overline{y})^2}{y_i}\right\}^2 \quad e$$

$$S_R^* = S_R\left\{1 - \frac{1}{18n}(33 - 34S_R + 4S_R^2)\right\};$$

Modelo gama $G(\mu,\phi)$:

$$S_R = \frac{-n\phi^{(0)}\{\log\phi^{(0)} - \psi_0 - \log(\overline{y}/\tilde{y})\}^2}{1-\phi^{(0)}\psi_0'}$$

e S_R^* *segue de (5.29) - (5.30) com*

$$A_1 = \frac{3}{n\phi^{(0)}(1-\phi^{(0)}\psi_0')}, \quad A_2 = \frac{3\phi^{(0)}(2\psi_0'' + \phi^{(0)}\psi_0''')}{n(1-\phi^{(0)}\psi_0')^2}$$

e

$$A_3 = \frac{-5(1+\phi^{(0)2}\psi_0'')^2}{n\phi^{(0)}(1-\phi^{(0)}\psi_0')^3},$$

onde $\psi_0, \psi_0', \ldots$ *denotam funções poligamas avaliadas em* $\phi = \phi^{(0)}$.

Exemplo 5.16 *Considera-se aqui o teste escore para o parâmetro* ρ *da distribuição exponencial tratada no exemplo 5.3. A estatística escore para testar* $H : \rho = \rho^{(0)}$ *versus* $A : \rho \neq \rho^{(0)}$ *é* $S_R = n(\rho^{(0)}\overline{y} - 1)^2$. *A estatística escore corrigida* S_R^* *segue facilmente de (5.32) como*

$$S_R^* = S_R\left\{1 - \frac{1}{18n}(3 - 11S_R + 2S_R^2)\right\}.$$

Os coeficientes desta estatística podem, também, ser calculados das equações (5.31), pois neste caso os momentos ordinários de S_R *até* $O(n^{-1})$ *são facilmente obtidos notando que* $n\overline{y}$ *tem distribuição gama com média* n/ρ *e índice igual a 1, a saber:* $\mu_1'(S_R) = 1$, $\mu_2'(S_R) = 3 + 4/n$ *e* $\mu_3'(S_R) = 15 + 130/n$. *Substituindo em (5.31) obtém-se os* $A_i's$ *e a mesma expressão de* S_R^* *dada anteriormente.*

Recentemente, vários artigos têm sido publicados apresentando as estatísticas escore corrigidas (5.29) em classes amplas de modelos de regressão. Cordeiro, Ferrari e Paula (1993) e Cribari-Neto e Ferrari (1995a) obtiveram correções tipo-Bartlett para testes escore em modelos lineares generalizados com parâmetro de dispersão conhecido e desconhecido, respectivamente. Correções similares para testes escore em modelos lineares heterocedásticos e em modelos não-lineares da família exponencial foram obtidos por Cribari-Neto e Ferrari (1995b) e Ferrari e Cordeiro (1996), respectivamente. No cálculo dessas correções tem sido mostrado através de estudos de simulação que as estatísticas escore modificadas por (5.29) são melhores aproximadas pela distribuição χ^2 de referência do que as estatísticas escore usuais. Uma revisão da literatura dos testes escore aperfeiçoados é dada por Cribari-Neto e Cordeiro (1996).

5.9 Aplicações à Família Exponencial

A família exponencial uniparamétrica, constitui um dos modelos estatísticos mais importantes, incluindo muitas distribuições clássicas. Além de um amplo espectro de aplicações, ela tem inúmeras propriedades interessantes (vide, por exemplo, Bickel e Doksum, 1977). O objetivo desta seção é apresentar o cálculo das correções de Bartlett para a razão de verossimilhança e tipo-Bartlett para a estatística escore na família exponencial especificada por um único parâmetro.

Considere um conjunto de n variáveis aleatórias *iid* com função densidade, ou no caso discreto com função de probabilidade, definida na família exponencial uniparamétrica

$$\pi(y;\theta) = \exp\{-\alpha(\theta)d(y) + v(y)\}/\zeta(\theta), \tag{5.33}$$

onde θ é um parâmetro escalar, $\zeta(\cdot)$, $\alpha(\cdot)$, $d(\cdot)$ e $v(\cdot)$ são funções conhecidas e $\zeta(\cdot)$ é positiva para todo θ no espaço de parâmetros. Admite-se que o conjunto suporte de (5.33) é independente de θ e que $\alpha(\cdot)$ e $\zeta(\cdot)$ têm derivadas contínuas até quarta ordem. Várias distribuições importantes em termos de aplicações à Economia, Engenharia, Biologia, Medicina, entre outras áreas, são membros da família (5.33), tais como as seguintes distribuições: geométrica, Bernoulli, binomial, binomial negativa, Poisson, Poisson truncada, série logaritmica, série de potências, zeta, hipergeométrica não-central, Maxwell, Erlang, exponencial, Rayleigh, Pareto, potência, valor extremo, valor extremo truncada, qui-quadrado e McCullagh (1989). Outras distribuições de dois parâmetros como normal, gama, log-normal, log-gama, Laplace e Weibull podem ser consideradas pertencentes à família exponencial (5.33) supondo que um de seus parâmetros é conhecido.

O objetivo aqui é corrigir as estatísticas da razão de verossimilhança e escore no teste de $H : \theta = \theta^{(0)}$ versus $A : \theta \neq \theta^{(0)}$, onde $\theta^{(0)}$ é um valor especificado para θ. Seja

$$\beta(\theta) = \{d\zeta(\theta)/d\theta\}/\{\zeta(\theta)d\alpha(\theta)/d\theta\}.$$

Verifica-se facilmente da função escore que $E\{-d(y)\} = \beta(\theta)$. A estimativa de máxima

verossimilhança $\hat{\theta}$ de θ é obtida iterativamente de $-n^{-1}\Sigma d(y_i) = \beta(\hat{\theta})$. As estatísticas w e S_R para o teste de H podem ser expressas por

$$w = 2n\beta(\hat{\theta})\{\alpha(\hat{\theta}) - \alpha(\theta^{(0)})\} + 2n\,\log\{\zeta(\theta^{(0)})/\zeta(\hat{\theta})\}$$

e $S_R = n\,d\alpha(\theta)/d\theta\ (\beta(\theta) + \overline{d})^2/(d\beta(\theta)/d\theta)$ com $\theta = \theta^{(0)}$, onde $\overline{d} = n^{-1}\Sigma d(y_i)$.

Seja $U(\theta) = -\alpha' d(y) - \zeta'/\zeta$ a função escore relativa a uma única observação. Derivadas em relação a θ são representadas por linhas. Observe-se que $E\{d(y)\} = -\beta(\theta)$. Sejam $v_r = v_r(\theta) = E\{U^{(r-1)}(\theta)\}$ e $v_{(r)} = v_{(r)}(\theta) = E\{U(\theta)^r\}$ para $r = 1, 2, 3$ e 4 e $v_{2(2)} = E\{U'(\theta)^2\}$. Os $v_i's$ estão relacionados com os cumulantes $\kappa's$ da Seção 5.2. Usando as identidades de Bartlett tem-se: $v_1 = 0, v_{(2)} = -v_2$, $v_{(3)} = 2v_3 - 3v_2'$ e $v_{(4)} = -3v_4 + 8v_3' - 6v_2'' + 3v_{2(2)}$. É fácil verificar através da função escore $U(\theta)$ que $v_2 = -\alpha'\beta'$, $v_3 = -2\alpha''\beta' - \alpha'\beta''$, $v_4 = -3(\alpha'''\beta' + \alpha''\beta'') - \alpha'\beta'''$ e $v_{2(2)} = \alpha''^2\beta'/\alpha' + \alpha'^2\beta'^2$.

Inserindo as equações acima na fórmula (5.23) obtém-se a correção de Bartlett para definir a razão de verossimilhança aperfeiçoada w^* no teste de $H : \theta = \theta^{(0)}$. Escreve-se esta correção como

$$c_B = 1 + \frac{\rho(\theta)}{12n}, \tag{5.34}$$

expressando a função $\rho(\theta)$ por (Cordeiro et al., 1995)

$$\rho(\theta) = \frac{-4\beta'^2\alpha''^2 - \alpha'\beta'\alpha''\beta'' + 5\alpha'^2\beta''^2 + 3\alpha'\beta'^2\alpha''' - 3\alpha'^2\beta'\beta'''}{\alpha'^3\beta'^3}. \tag{5.35}$$

A fórmula de $\rho(\theta)$ em (5.35) depende apenas de α e β e de suas três primeiras derivadas em relação a θ. Quando α é linear em θ, correspondente à família exponencial natural, tem-se a redução simples $\rho(\theta) = (5\beta''^2 - 3\beta'\beta'')/\beta'^3$. Pode-se, então, calcular a correção para w em qualquer distribuição de (5.33) inserindo simplesmente as funções correspondentes α e β, e suas derivadas, na equação (5.35). Uma dificuldade na interpretação desta equação é que os termos individuais não são invariantes em relação a reparametrização e, portanto, eles não têm interpretação geométrica independente do sistema de coordenadas especificado.

Deduz-se, agora, o teste escore aperfeiçoado da hipótese $H : \theta = \theta^{(0)}$. A partir da equação (5.32) e usando as diversas relações entre os $v_i's$, obtêm-se, após extensiva álgebra, os coeficientes $a = a(\theta)$, $b = b(\theta)$ e $c = c(\theta)$ da estatística escore modificada $S_R^* = S_R\{1 - \frac{1}{n}(c + bS_R + aS_R^2)\}$, deduzidas por Ferrari et al. (1996) como

$$a = \frac{(\beta'\alpha'' - \alpha'\beta'')^2}{36\alpha'^3\beta'^3}$$

e (5.36)

$$b = \frac{-\beta'^2\alpha''^2 + 11\alpha'\beta'\alpha''\beta'' - 10\alpha'^2\beta''^2 - 3\alpha'\beta'^2\alpha''' + 3\alpha'^2\beta'\beta'''}{36\alpha'^3\beta'^3}.$$

O coeficiente c segue, diretamente, de (5.35) como $c = \rho(\theta)/12$. Substituindo as equações (5.36) e c na fórmula de S_R^*, obtém-se a estatística escore melhorada para testar $H : \theta = \theta^{(0)}$ na família exponencial uniparamétrica. Os coeficientes a e b, a exemplo de $\rho(\theta)$, dependem do modelo apenas através das funções α e β e de suas três primeiras derivadas. A grande vantagem das equações (5.35) – (5.36) é que elas não requerem o cálculo de cumulantes mas somente das derivadas de α e β. Claramente, a, b e c são avaliados em $\theta = \theta^{(0)}$ para calcular numericamente as correções.

Da equação (5.35) pode-se demonstrar que $\rho(\theta) = 2$ se: (i) $\alpha(\theta)\zeta(\theta) = c_1$, ou (ii) $\alpha(\theta)$ é linear, por exemplo $\alpha(\theta) = c_1\theta + c_2$, e $\zeta(\theta) = c_3/\theta c_4^\theta$, onde c_1, c_2, c_3 e c_4 são constantes arbitrárias. Estas condições são individualmente suficientes, mas não são necessárias, para garantir $\rho(\theta) = 2$. Também, demonstra-se, das equações (5.36) que as condições (i) e (ii) são, também, individualmente suficientes para que se tenha $a = 1/9$ e $b = -11/18$, implicando, então, que a estatística escore modificada seja da forma $S_R^* = S_R\{1 - (3 - 11S_R + 2S_R^2)/(18n)\}$. Pode-se verificar que isto ocorre para várias distribuições que satisfazem uma das condições acima.

As equações (5.35) e (5.36) são, facilmente, calculadas algebricamente com o auxílio de programas de computação simbólica como REDUCE, MATHEMATICA (Wolfram, 1996) e MAPLE. Cordeiro et al. (1995) e Ferrari et al. (1996) apresentam, repectivamente, fórmulas especiais para $\rho(\theta)$, $a(\theta)$, $b(\theta)$ e $c(\theta)$ em 30 distribuições da família exponencial (5.33). Seguem abaixo, oito exemplos, onde $\rho(\theta) = 12c(\theta)$:

(i) Binomial ($0 < \theta < 1$, $m \in \mathbb{N}$, m conhecido, $y = 0, 1, 2, \ldots, m$): $\alpha(\theta) = -\log\{\theta/(1-\theta)\}$, $\zeta(\theta) = (1-\theta)^{-m}$, $d(y) = y$, $v(y) = \log\binom{m}{y}$:

$$a = \frac{(2\theta-1)^2}{36m\theta(1-\theta)}, \quad b = \frac{22\theta(\theta-1)+7}{36m\theta(\theta-1)}, \quad c = \frac{\theta(1-\theta)-1}{6m\theta(\theta-1)}.$$

(ii) Poisson ($\theta > 0$, $y = 0, 1, 2, \ldots$): $\alpha(\theta) = -\log\theta$, $\zeta(\theta) = \exp(\theta)$, $d(y) = y$, $v(y) = -\log y!$: $a = 1/(36\theta)$, $b = -7/(36\theta)$, $c = 1/(6\theta)$.

(iii) Normal ($\theta > 0$, $-\infty < \mu < \infty$, $-\infty < y < \infty$):

(a) μ conhecido: $\alpha(\theta) = (2\theta)^{-1}$, $\zeta(\theta) = \theta^{1/2}$, $d(y) = (y-\mu)^2$, $v(y) = -\{\log(2\pi)\}/2$: $a = 2/9$, $b = -11/9$, $c = 1/3$.

(b) θ conhecido: $\alpha(\mu) = -\mu/\theta$, $\zeta(\mu) = \exp\{\mu^2/(2\theta)\}$, $d(y) = y$, $v(y) = -\{y^2 + \log(2\pi\theta)\}/2$: $a = 0$, $b = 0$, $c = 0$.

(iv) Normal Inversa ($\theta > 0$, $\mu > 0$, $y > 0$):

(a) μ conhecido: $\alpha(\theta) = \theta$, $\zeta(\theta) = \theta^{-1/2}$, $d(y) = (y-\mu)^2/(2\mu^2 y)$, $v(y) = -\{\log(2\pi y^3)\}/2$: $a = 2/9$, $b = -11/9$, $c = 1/3$.

(b) θ conhecido: $\alpha(\mu) = \theta/(2\mu^2)$, $\zeta(\mu) = \exp(-\theta/\mu)$, $d(y) = y$, $v(y) = -\theta/(2y) + [\log\{\theta/(2\pi y^3)\}]/2$: $a = \mu/(4\theta)$, $b = -5\mu/(4\theta)$, $c = 0$.

(v) Gama ($k > 0, \theta > 0$, $y > 0$):

(a) k conhecido: $\alpha(\theta) = \theta$, $\zeta(\theta) = \theta^{-k}$, $d(y) = y$, $v(y) = (k-1)\log y - \log\{\Gamma(k)\}$: $a = 1/(9k)$, $b = -11/(18k)$, $c = 1/(6k)$.

(b) θ conhecido: $\alpha(k) = 1 - k$, $\zeta(k) = \theta^{-k}\Gamma(k)$, $d(y) = \log y$, $v(y) = -\theta y$:

$$a = \frac{\psi''(k)^2}{36\psi'(k)^3}, \quad b = \frac{-10\psi''(k)^2 + 3\psi'(k)\psi'''(k)}{36\psi'(k)^3}, \quad c = \frac{5\psi''(k)^2 - 3\psi'(k)\psi'''(k)}{12\psi'(k)^3},$$

onde $\psi(\cdot)$ é a função digama.

(vi) Rayleigh ($\theta > 0$, $y > 0$): $\alpha(\theta) = \theta^{-2}$, $\zeta(\theta) = \theta^2$, $d(y) = y^2$, $v(y) = \log(2y)$: $a = 1/9$, $b = -11/18$, $c = 1/6$.

(vii) Pareto ($\theta > 0$, $k > 0$, k conhecido, $y > k$): $\alpha(\theta) = \theta + 1$, $\zeta(\theta) = (\theta k^\theta)^{-1}$, $d(y) = \log y$, $v(y) = 0$: $a = 1/9$, $b = -11/18$, $c = 1/6$.

(viii) Weibull ($\theta > 0$, $\phi > 0$, ϕ conhecido, $y > 0$): $\alpha(\theta) = \theta^{-\phi}$, $\zeta(\theta) = \theta^{\phi}$, $d(y) = y^{\phi}$, $v(y) = \log\phi + (\phi-1)\log y$: $a = 1/9$, $b = -11/18$, $c = 1/6$.

5.10 Exercícios

1. Seja Y_n uma variável aleatória que tende em distribuição para uma variável χ^2_q quando $n \longrightarrow \infty$. Seja $f_q(y)$ a função densidade de χ^2_q. Demonstre que as expansões seguintes para as funções densidade $f_n(y)$ e geratriz de momentos $M_n(t)$ de Y_n são equivalentes até $O(n^{-1})$:

   (a) $f_n(y) = f_q(y)(1 - \frac{c}{n}) + f_{q+2}(y)\frac{c}{n}$;
   (b) $f_n(y) = f_q(y)\{1 + \frac{c}{n}(\frac{y}{q} - 1)\}$;
   (c) $M_n(t) = (1 - 2t)^{-q/2}\{1 + \frac{2ct}{n}(1 - 2t)^{-1}\}$.

   Mostre, também, que a função densidade de $Y_n(1 + \frac{2c}{nq})$ é $f_q(y)$ com erro $o(n^{-1})$ e, portanto, $(1 + \frac{2c}{nq})$ é a correção de Bartlett de Y_n.

2. Demonstre que o viés de ordem n^{-1} da EMV $\hat{\theta}$ do parâmetro θ da família exponencial (5.33) é deduzido da equação (5.5) como $\beta(\hat{\theta}) = -\beta''/(2n\alpha'\beta'^2)$, onde $\beta(\theta) = -E\{d(y)\}$ e as derivadas são em relação ao parâmetro θ.

3. Para as distribuições (i) – (viii) da família exponencial (5.33) apresentadas na Seção 5.9 mostre que o viés $B(\theta)$ da EMV $\hat{\theta}$, obtido da equação (5.5), é dado por:

   (i) Binomial: $B(\hat{\theta}) = 0$;
   (ii) Poisson: $B(\hat{\theta}) = 0$;
   (iii) Normal: (a) μ conhecido, $B(\hat{\theta}) = 0$; (b) θ conhecido, $B(\hat{\mu}) = 0$;
   (iv) Normal Inversa: (a) μ conhecido, $B(\hat{\theta}) = 2\theta/n$; (b) θ conhecido, $B(\hat{\mu}) = 0$;
   (v) Gama: (a) k conhecido, $B(\hat{\theta}) = \theta/(nk)$; (b) θ conhecido, $B(\hat{\theta}) = \frac{-\psi''(k)}{2n\psi'(k)^2}$, onde $\psi(\cdot)$ é a função digama;
   (vi) Rayleigh: $B(\hat{\theta}) = -\theta/(8n)$;
   (vii) Pareto: $B(\hat{\theta}) = \theta/n$;
   (viii) Weibull: $B(\hat{\theta}) = \theta(1 - \phi)/(2n\phi^2)$.

4. Suponha a distribuição χ^2_θ com número de graus de liberdade θ desconhecido. (a) Calcule a aproximação de Barndorff-Nielsen para a função densidade da EMV $\hat{\theta}$; (b) Calcule o viés de ordem n^{-1} de $\hat{\theta}$; (c) Calcule as correções de Bartlett e tipo-Bartlett para melhorar as estatísticas da razão de verossimilhança e escore no teste de $H : \theta = \theta^{(0)}$ versus $A : \theta \neq \theta^{(0)}$.

5. Usando (5.10) calcule a aproximação para a função densidade da EMV $\hat{\theta}$ nas distribuições (i) - (viii) da família exponencial (5.33) descritas na Seção 5.9.

6. A distribuição de von Mises usada para análise de dados circulares é um membro da família exponencial (5.33) onde ($\theta > 0$, $0 < \mu < 2\pi, \mu$ conhecido, $0 < y < 2\pi$): $\alpha(\theta) = \theta$, $\zeta(\theta) = 2\pi I_0(\theta)$, $d(y) = \cos(y - \mu)$, $v(y) = 0$ e $I_\nu(\cdot)$ é a função de Bessel de primeira espécie e ordem ν. (a) Determine o viés de ordem n^{-1} da EMV $\hat{\theta}$; (b) Das equações (5.35) - (5.36) encontre as correções de Bartlett e tipo-Bartlett para melhorar as estatísticas da razão verossimilhança e escore no teste de $H : \theta = \theta^{(0)}$ versus $A : \theta \neq \theta^{(0)}$; (b) Deduza de (5.10) a aproximação para a função densidade de $\hat{\theta}$.

7. Para os modelos log-gama, gama e logístico descritos nos exemplos 5.10 e 5.11, apresente fórmulas para aproximar $P(Y \leq y; \theta)$ baseadas em (5.15).

8. Caracterize as seguintes distribuições de um parâmetro: geométrica, binomial negativa, Poisson truncada, série logaritmica, série de potências, Maxwell, Pareto, Rayleigh, valor extremo, lognormal e potência, como membros da família exponencial (5.33). (a) Deduza das equações (5.35) - (5.36) as correções para melhorar os testes da razão de verossimilhança e escore (Cordeiro et al., 1995; Ferrari et al., 1996). (b) Deduza fórmulas para os vieses de ordem n^{-1} das EMV do parâmetro que caracteriza estas distribuições.

9. Sejam n observações independentes $y_1, \ldots, y_n$ de uma distribuição de Poisson com a estrutura log linear $\log \mu_i = \alpha + \beta x_i$, $i = 1, \ldots, n$. Suponha o teste de $H : \beta = 0$ versus $A : \beta \neq 0$. Demonstre que a estatística escore para este teste é $S_R = n\overline{s}^2(\overline{y}\ \overline{s}_2)^{-1}$ e que A_1, A_2 e A_3 são obtidos das equações (5.28) como: $A_1 = 0$, $A_2 = -3(3 - \overline{s}_4/\overline{s}_2^2)(n\tilde{\mu})^{-2}$ e $A_3 = 5\overline{s}_3^2/(n\tilde{\mu}\overline{s}_2^3)$, onde $\overline{s}_a = \sum_{i=1}^{n}(x_i - \overline{x})^a/n$ e $\tilde{\mu} = \overline{y}$.

Referências

Attfield, C.L.F. (1991). *A Bartlett-adjustment to the likelihood ratio test for homoskedasticity in the linear model.* Economics Letters, **37**, 119–123.

Barndorff-Nielsen, O.E. (1983). *On a formula for the distribution of the maximum likelihood estimator.* Biometrika, **70**, 343–365.

Barndorff-Nielsen, O.E. (1986). *Inference on full and partial parameters based on the standardized signed log-likelihood ratio.* Biometrika, **73**, 307–322.

Barndorff-Nielsen, O.E. (1988). *Contribution to discussion of paper by N. Reid (1988).* Statistical Science, **3**, 228–229.

Barndorff-Nielsen, O.E. (1990). *Approximate interval probabilities.* J.R. Statist. Soc. B, **52**, 485–496.

Barndorff-Nielsen, O.E. e Blaesild, P. (1986). *A note on the calculation of Bartlett adjustments.* J. R. Statist. Soc. B, **46**, 483–495.

Barndorff-Nielsen, O.E. e Cox, D.R. (1984a). *Bartlett adjustments to the likelihood ratio statistic and the distribution of the maximum likelihood estimator.* J. R. Statist. Soc. B, **46**, 484–495.

Barndorff-Nielsen, O.E. e Cox, D.R. (1984b). *The effect of sampling rules on likelihood statistics.* Int. Statist. Rev., **52**, 309–326.

Barndorff-Nielsen, O.E. e Cox, D.R. (1990). *Asymptotic Techniques for use in Statistics.* Londres: Chapman and Hall.

Barndorff-Nielsen, O.E. e Cox, D.R. (1994). *Inference and Asymptotics.* Londres: Chapman and Hall.

Barndorff-Nielsen, O.E. e Hall, P. (1988). *On the level-error after Bartlett adjusment of the likelihood ratio statistic.* Biometrika, **75**, 374–378.

Bartlett, M.S. (1953). *Approximate Confidence Intervals I.* Biometrika, **40**, 12–19.

Bickel, P.J. e Doksum, K.A. (1977). *Mathematical Statistics: Basic Ideas and Selected Topics.* Oakland: Holden-Day.

Bleistein, N. e Handelsman, R.A. (1975). *Asymptotic Expansions of Integrals.* Nova York: Holt, Rinehart and Winston.

Breusch, T.S. e Pagan, A.R. (1980). *The Lagrange multiplier test and its applications to model specification in econometrics.* Review of Economic Studies, **47**, 239–253.

Chandra, T.K. (1985). *Asymptotic expansions of perturbed chi-square variables.* Sankhyã A, **47**, 100–110.

Chesher, A. e Smith, R.J. (1995). *Bartlett corrections to likelihood ratio tests.* Biometrika, **82**, 433–436.

Cordeiro, G.M. (1983). *Improved likelihood ratio statistics for generalized linear models.* J. R. Statist. Soc. B, **45**, 404–413.

Cordeiro, G.M. (1987). *On the corrections to the likelihood ratio statistics.* Biometrika, **74**, 265–274.

Cordeiro, G.M. (1993). *Bartlett corrections and bias correction for two heteroscedastic regression models.* Comm. Statist. Theor. and Meth., **22**, 169–188.

Cordeiro, G.M. e Cribari-Neto, F. (1998). *On Bias reduction in exponential and non-exponential family regression models.* Comm. Statist. Simul. Comput., **27**, 485–500.

Cordeiro, G.M., Cribari-Neto, F., Aubin, E.C.Q. e Ferrari, S.L.P. (1995). *Bartlett corrections for one-parameter exponential family models.* J. Statist. Comput. Simul., **53**, 211–231.

Cordeiro, G.M. e Ferrari, S.L.P. (1991). *A modified score statistic having chi-squared distribution to order n^{-1}.* Biometrika, **78**, 573–582.

Cordeiro, G.M., Ferrari, S.L.P. e Paula, G.A. (1993). *Improved score tests for generalized linear models.* J. R. Statist. Soc. B, **55**, 661–674.

Cordeiro, G.M. e Ferrari, S.L.P. (1998). *A note on Bartlett-type corrections for the first few moments of test statistics.* J. Statist. Plan. Infer., **71**, 261–269.

Cordeiro, G.M. e Klein, R. (1994). *Bias correction in ARMA models.* Statist. Probab. Lett., **19**, 169–176.

Cordeiro, G.M. e McCullagh, P. (1991). *Bias correction in generalized linear models.* J.R. Statist. Soc. B, **53**, 629–643.

Cordeiro, G.M. e Paula, G.A. (1989). *Improved likelihood ratio statistics for exponential family nonlinear models.* Biometrika, **76**, 93–100.

Cordeiro, G.M., Paula, G.A. e Botter, D.A. (1994). *Improved likelihood ratio tests for dispersion models.* Int. Statist. Rev., **62**, 257–276.

Cox, D.R. e Hinkley, D.V. (1979). *Theoretical Statistics.* Nova York: Chapman and Hall.

Cox, D.R. e Reid, N. (1987). *Parameter orthogonality and approximate conditional inference (with discussion).* J.R. Statist. Soc. B, **49**, 1–39.

Cox, D.R. e Snell, E.J. (1968). *A general definition of residuals (with discussion).* J.R. Statist. Soc. B, **30**, 248–278.

Cramér, H. (1937). *Random Variables and Probability Distributions.* Londres: Cambridge University Press.

Cramér, H. (1946). *Mathematical Methods of Statistics.* Princeton: Princeton University Press.

Cribari-Neto, F. e Cordeiro, G.M. (1996). *On Bartlett and Bartlett-type corrections.* Econometric Reviews, **15**, 339–367.

Cribari-Neto, F. e Ferrari, S.L.P. (1995a). *Second order asymptotics for score tests in generalized linear models.* Biometrika, **82**, 426–432.

Cribari-Neto, F. e Ferrari, S.L.P. (1995b). *Bartlett-corrected tests for heteroskedastic linear models.* Econometric Letters, **48**, 113-118.

Daniels, H.E. (1954). *Saddlepoint approximations in Statistics.* Ann. Math. Statist. **25**, 631-650.

Daniels, H.E. (1983). *Saddlepoint approximations for estimating equations.* Biometrika, **70**, 89-96.

Daniels, H.E. (1987). *Tail probability approximations.* Int. Stat. Rev., **55**, 37–48.

Davison, A.C. e Hinkley, D.V. (1988). *Saddlepoint approximations in resampling methods.* Biometrika, **75**, 417–432.

DeBruijn, N.G. (1970). *Asymptotic Methods in Analysis.* Amsterdam: North-Holland.

DiCiccio, T.J., Field, C.A. e Fraser, D.A.S. (1990). *Approximation of marginal tail probabilities and inference for scalar parameters.* Biometrika, **77**, 77–95.

Edwards, A.W.F. (1972). *Likelihood.* Cambridge: Cambridge University Press.

Engle, R.F. (1984). *Wald, likelihood ratio and Lagrange multiplier tests in econometrics.* In: Griliches, Z. & Intriligator, M.(eds). Handbook of Econometrics. Amsterdam: North-Holland.

Feller, W. (1971). *An Introduction to Probability Theory and Its Applications.* Nova York: John Wiley.

Ferrari, S.L.P., Botter, D.A., Cordeiro, G.M. e Cribari-Neto, F. (1996). *Second and third order bias reduction in one-parameter models.* Statist. Probab. Lett., **30**, 339–345.

Ferrari, S.L.P. e Cordeiro, G.M. (1994). *Matrix formulae for computing improved score tests.* J. Statist. Comput. Simul., **49**, 196–206.

Ferrari, S.L.P. e Cordeiro, G.M. (1996). *Corrected score tests for exponential family nonlinear models.* Statist. Probab. Lett., **26**, 7–12.

Ferrari, S.L. de P., Cordeiro, G.M., Uribe-Opazo, M.A. e Cribari-Neto, F. (1996). *Improved score tests for one-parameter exponential family models.* Statist. Probab. Lett., **30**, 61–71.

Fisz, M. (1963). *Probability Theory and Mathematical Statistics.* Nova York: John Wiley.

Fraser, D.A.S. (1968). *The Structure of Inference.* Nova York: John Wiley.

Fraser, D.A.S. (1988). *Normed likelihood as saddlepoint approximation.* J. Mult. Anal., **27**, 181–193.

Fraser, D.A.S. (1990). *Tail probabilities from observed likelihoods.* Biometrika, **77**, 65–76.

Harris, P. (1985). *An asymptotic expansion for the null distribution of the efficient score statistic.* Biometrika, **72**, 653–659.

Hayakawa, T. (1977). *The likelihood ratio criterion and the asymptotic expansion of its distribution.* Ann. Inst. Statist. Math. A, **29**, 359–378.

Hinkley, D.V., Reid, N. e Snell, E.J.(eds) (1991). *Statistical Theory and Modelling.* Londres: Chapman and Hall.

Hosking, J.R.M. (1980). *Lagrange multiplier tests of time-series model.* J.R. Statist. Soc. B, **42**, 170–181.

Hosking, J.R.M. (1981). *Lagrange multiplier tests of multivariate time-series models.* J.R. Statist. Soc. B, **43**, 219–230.

Ibragimov, I.A. e Linnik, Yu.V. (1971). *Independent and Stationary Sequences of Random Variables.* Groningen: Wolters-Noordhoff.

Jeffreys, H. (1962). *Asymptotic Approximations.* Oxford: Oxford University Press.

Jensen, J.L. (1988). *Uniform saddlepoint approximations.* Adv. Appl. Prob., **20**, 622–634.

Jørgensen, B. (1987). *Exponential dispersion models (with discussion).* J.R. Statist. Soc. B, **49**, 127–162.

Kendall, e Rao, K.S. (1950). *On the generalized second limit theorem in the theory of probabilities.* Biometrika, **37**, 224.

Kolassa, J. e McCullagh, P. (1990). *Edgeworth expansions for discrete distributions.* Ann. Statist., **18**, 981–985.

Lawley, D.N. (1956). *A general method for approximating to the distribution of the likelihood ratio criteria.* Biometrika, **71**, 233–244.

Lawless, J.F. (1982). *Statistical Models and Methods for Lifetime Data.* Nova York: John Wiley.

LeCam, L. (1956). *On the asymptotic theory of estimation and testing hypotheses.* Proc. 3rd Berkeley Symp., **1**, 129–156.

LeCam, L. (1970). *On the assumptions used to prove asymptotic normality of maximum likelihood estimates.* Ann. Math. Statist., **41**, 802–828.

Lehmann, E.L. (1959). *Testing Statistical Methods.* Nova York: John Wiley.

Lehmann, E.L. (1983). *Theory of Point Estimation.* Nova York: John Wiley.

Lehmann, E.L. e Scheffé, H. (1950). *Completeness, similar regions, and unbiased estimation, Part I.* Sankhyã, **10**, 305–340.

Lévy, P. (1937). *Theorie de l'addition des variables aléatoires.* Paris: Gauthier-Villars.

Lugannani, R. e Rice, S. (1980). *Saddlepoint approximation for the distribution of the sum of independent random variables.* Adv. Appl. Prob., **12**, 475–490.

McCullagh, P. (1984). *Tensor notation and cumulants of polynomials.* Biometrika, **71**, 461–476.

McCullagh, P. (1987). *Tensor methods in Statistics.* Londres: Chapman and Hall.

McCullagh, P. (1989). *Some statistical properties of a family of continuous univariate distributions.* J. Amer. Statist. Assoc., **84**, 125–141.

McCullagh, P. e Cox, D.R. (1986). *Invariants and likelihood ratio statistics.* Ann. Statist., **14**, 1419–1430.

Møller, J. (1986). *Bartlett adjustments for structured covariances.* Scand. J. Statist., **13**, 1–15.

Paula, G.A. e Cordeiro, G.M. (1995). *Bias correction and improved residuals for non-exponential family nonlinear models.* Comm. Statist. Simul. Comput., **24**, 1193–1210.

Porteous, B.T. (1985). *Improved likelihood ratio statistics for covariance selection models.* Biometrika, **72**, 97–101.

Poskitt, D.S. e Tremayne, A.R. (1981). *An approach to testing linear time series models.* Ann. Statist. 9, 974–986.

Poskitt, D.S. e Tremayne, A.R. (1982). *Diagnostic tests for multiple time series models.* Ann. Statist., **10**, 114–120.

Pratt, J.W. (1968). *A normal approximation for binomial, F, beta and other common related tail probabilities.* J. Amer. Statist. Assoc., **63**, 1457–1483.

Pregibon, D. (1982). *Score tests in GLIM with applications.* Lecture Notes in Statistics, **14**, 87–97.

Rao, C.R. (1973). *Linear Statistical Inference and Its Applications.* Nova York: John Wiley.

Reid, N. (1988). *Saddlepoint methods and statistical inference.* Statistical Science, **3**, 213–238.

Ross, W.H. (1987). *The expectation of the likelihood ratio criterion.* Int. Statist. Rev., **55**, 315–330.

Serfling, R.J. (1980). *Approximation Theorems of Mathematical Statistics.* Nova York: John Wiley.

Shenton, L.R. e Bowman, K.O. (1977). *Maximum likelihood Estimation in Small Samples.* Londres: Charles Griffin.

Wilks, S.S. (1938). *The large-sample distribution of the likelihood ratio for testing composite hypotheses.* Ann. Math. Statist., **9**, 60–62.

Wilks, S.S. (1962). *Mathematical Statistics.* Nova York: John Wiley.

Wolfram, S. (1996). *The Mathematica Book.* New York: Addison-Wesley.

Impresso na Gráfica do

Impa

pelo Sistema Xerox / 5390