

Bootstrap e subamostragem em dados censurados tipo I

João Renato Sebastião

Escola Superior de Gestão do Instituto Politécnico de Castelo Branco, Departamento de Matemática e Informática

Sara Morgado Nunes

Escola Superior de Gestão do Instituto Politécnico de Castelo Branco, Departamento de Matemática e Informática

Resumo: A censura está, em geral, presente nos Testes de Vida. Quando a censura em causa é do Tipo I, não existem métodos analíticos que possibilitem a construção de intervalos de confiança, dificuldade esta que se deve à complexidade da respectiva função de verosimilhança. Neste trabalho consideram-se o bootstrap e a subamostragem, enquanto procedimentos úteis na determinação de intervalos de confiança para os parâmetros da distribuição do tempo de vida, num contexto de censura Tipo I. Num estudo de simulação aplicam-se estas metodologias a amostras censuradas artificialmente, provenientes de uma distribuição Weibull, com o objectivo de determinar intervalos de confiança para o quantil de probabilidade p .

Palavras-chave: Censura tipo I, Intervalo de confiança, Bootstrap e subamostragem

Abstract: Censoring is, in general, present in Life Tests. When data are Type I censored, analytic methods don't exist for the construction of confidence intervals. This difficulty is due to the likelihood function complexity. This paper presents bootstrap and subsampling as useful tools to obtain confidence intervals for the parameters in Type I censoring context. In a simulation study these methodologies are applied to artificially censored samples in order to determine confidence intervals for the p quantile.

Keywords: Type I censoring, Confidence interval, Bootstrap and subsampling

1 Introdução

O desenvolvimento tecnológico da sociedade actual impõe, cada vez mais, a implementação de medidas que possibilitem o melhoramento dos níveis de produção e prestação de serviços. Desta forma, tem surgido um interesse crescente na literatura científica pelos Testes de Vida (Meeker e Escobar, 1998), nos quais, em geral, está presente a censura. Apesar de ser empregue desde longa data, a Fiabilidade passa a ser alvo de grande interesse a partir da I Guerra Mundial em que era necessário comparar a segurança operacional de aviões bombardeiros de um, dois ou quatro motores, medindo-se o número de acidentes por hora de voo. Desde então, as técnicas de fiabilidade, aliadas ao desenvolvimento tecnológico, registaram um progresso extraordinário e têm hoje aplicação nos mais diver-

nos sectores industriais desde estudos de materiais, análise de riscos, protecção ambiental, medicina, até à tecnologia espacial como a construção de mísseis e centrais nucleares. Apesar de se conhecerem várias definições de Fiabilidade, esta surge sempre associada à capacidade física de um dado sistema cumprir uma determinada função específica, isto é, funcionar correctamente durante o período de tempo para o qual foi concebido.

No estudo de tempos de vida, podem ocorrer dados censurados por razões variadas. Uma observação relativa a uma variável aleatória T diz-se censurada à direita quando se sabe unicamente que o seu possível tempo de falha ocorreu para além de um determinado valor l_i , isto é, a eventual falha ocorreu no intervalo $[l_i, +\infty[$. Por outro lado, uma observação diz-se censurada à esquerda quando se sabe que a falha ocorreu antes de um determinado instante l_s , isto é, a falha ocorreu no intervalo $[0, l_s]$. Se o teste termina após um período de tempo pré-determinado, a censura em causa diz-se Tipo I ou censura pelo tempo, enquanto a censura Tipo II ou censura pela falha, está presente quando o teste termina após ter falhado um número pré-estabelecido de unidades.

Seja $Y = \{(t_1, \delta_1), (t_2, \delta_2), \dots, (t_n, \delta_n)\}$ uma amostra aleatória de dados censurados à direita, em que

$$\delta_i = \begin{cases} 1 & \text{, se } t_i \text{ é observação não censurada} \\ 0 & \text{, se } t_i \text{ é observação censurada} \end{cases}, \quad i = 1, \dots, n.$$

Admitindo que a variável T possui distribuição pertencente à família log-localização escala, a função de distribuição é dada por

$$F(t; \mu, \sigma) = \Phi\left(\frac{\log(t) - \mu}{\sigma}\right), \quad t > 0, \mu \in \mathbb{R} \text{ e } \sigma > 0 \quad (1)$$

e a função densidade de probabilidade por

$$f(t; \mu, \sigma) = \frac{1}{\sigma t} \phi\left(\frac{\log(t) - \mu}{\sigma}\right), \quad t > 0, \mu \in \mathbb{R} \text{ e } \sigma > 0 \quad (2)$$

onde, Φ e ϕ são funções que dependem da distribuição em causa mas não dos parâmetros de localização μ e escala σ . Por exemplo, se a distribuição em causa for a Weibull, Φ e ϕ definem-se por

$$\Phi(z) = 1 - \exp(-\exp(z)) \quad (3)$$

$$\phi(z) = \exp(z - \exp(z)). \quad (4)$$

Suponha-se que se opta por uma estatística $\hat{\theta}$ para estimar um determinado parâmetro populacional θ . No caso de $\hat{\theta}[(t_1, \delta_1), (t_2, \delta_2), \dots, (t_n, \delta_n)]$ ser o estimador de máxima verosimilhança de θ com base na amostra Y , o problema que se coloca é o de avaliar a precisão de $\hat{\theta}$ enquanto estimador de θ . Em contextos de censura Tipo I não existem métodos analíticos que permitam calcular

intervalos de confiança exactos, uma vez que a função de verosimilhança contém uma componente discreta, relacionada com o número aleatório de falhas, que dificulta a determinação de estatísticas suficientes, necessárias ao procedimento de estimação exacta.

Foi justamente numa tentativa de avaliar a precisão de uma estatística $\hat{\theta}$ enquanto estimador de θ e de estimar o seu erro padrão que surgiram os métodos de reamostragem como o bootstrap [Efron (1979)] e mais tarde a subamostragem de Politis e Romano (1994). A ideia subjacente a estas metodologias é a utilização repetida da amostra original de forma a obter várias estimativas de θ que serão usadas para aproximar a função de distribuição de $\hat{\theta}$ e as suas características.

Nas secções 2 e 3 procede-se a uma revisão muito breve das metodologias bootstrap e subamostragem enquanto ferramentas úteis na construção de intervalos de confiança para um parâmetro associado a uma determinada distribuição de probabilidade, num contexto de dados censurados. Na secção 4 apresentam-se os resultados de um estudo de simulação em que se aplicam as referidas metodologias a amostras censuradas artificialmente, geradas por uma distribuição Weibull, com o objectivo de calcular intervalos de confiança para o quantil de probabilidade p .

2 Bootstrap (PBSRLR)

O bootstrap faz parte da grande família dos métodos de reamostragem e foi originalmente proposto por Efron em 1979, tendo revolucionado a Estatística. Embora as ideias subjacentes ao bootstrap sejam antigas, o interesse da comunidade científica pelo método surgiu essencialmente após os trabalhos de Efron (1979) e Efron e Tibshirani (1993) onde o bootstrap é apresentado como um método de base computacional adequado a vários problemas de estimação. Desde então, o bootstrap tem vindo a ser aperfeiçoado, constituindo actualmente um método computacionalmente intensivo eficaz e amplamente usado na Inferência Estatística.

Com o objectivo de dar resposta ao problema apresentado, o bootstrap permite construir, de modo simples, estimadores para a distribuição de $\hat{\theta}$ com base na amostra Y . Assim, começa-se por construir a distribuição empírica \hat{F} , estimador de F , a qual se obtém atribuindo peso $\frac{1}{n}$ a cada um dos valores observados (t_i, δ_i) , $i = 1, \dots, n$. Com \hat{F} fixo, defina-se uma amostra bootstrap $Y^* = \{(t_1, \delta_1)^*, (t_2, \delta_2)^*, \dots, (t_n, \delta_n)^*\}$ como sendo uma amostra aleatória de dimensão n obtida de \hat{F} com reposição a partir de Y . A notação $*$ indica que Y^* não é o conjunto de dados original Y mas sim uma versão reamostrada de Y . Assim, seleccionam-se B amostras bootstrap independentes de dimensão n , $Y^{*1}, Y^{*2}, \dots, Y^{*B}$, obtidas com reposição a partir de Y . Associada a cada conjunto de dados bootstrap Y^* está uma réplica bootstrap de $\hat{\theta}$ notada por $\hat{\theta}_b^*$, $b = 1, \dots, B$, a qual resulta do cálculo da estatística de interesse em Y^* . A

partir das B réplicas bootstrap é possível obter vários intervalos de confiança (Davison e Hinkley, 1997) para θ . Porém, num contexto de dados censurados, é comum recorrer-se à metodologia PBSRLLR (Parametric Bootstrap Signed Square Root Log Likelihood Ratio) que consiste em utilizar a simulação bootstrap para obter a distribuição amostral da estatística SRLLR em vez de recorrer a uma aproximação à distribuição Normal. De facto, verifica-se que os intervalos de confiança PBSRLLR conduzem a melhores aproximações que intervalos de confiança baseados em aproximações à distribuição Normal em contextos de censura Tipo I (Ostrouchov e Meeker, 1988). Vander Wiel e Meeker (1990) mostram ainda que estes intervalos de confiança apresentam probabilidades de cobertura mais próximas da probabilidade nominal que os obtidos através dos procedimentos de aproximação à distribuição Normal.

Seja $\theta = (\theta_1, \theta_2)$ um vector de parâmetros desconhecidos onde θ_1 é o parâmetro de interesse e θ_2 o parâmetro perturbador. Defina-se a função "Profile Likelihood" para θ_1 como

$$R(\theta_1) = \max_{\theta_2} \left[\frac{L(\theta_1, \theta_2)}{L(\hat{\theta})} \right], \quad (5)$$

onde L denota a função de verosimilhança. Seja

$$V(\theta_1) = \text{sign}(\hat{\theta}_1 - \theta_1) [-2 \log R(\theta_1)]^{\frac{1}{2}} \quad (6)$$

e $v_{\theta_1}^*(\alpha)$ o *quantil* $-\alpha$ da distribuição bootstrap $V(\theta_1)$. Os limites inferior e superior do **intervalo de confiança PBSRLLR** de $(1 - \alpha)100\%$ para θ_1 são dados, respectivamente, pelas raízes de

$$V(\theta_1) - v_{\theta_1}^*\left(\frac{\alpha}{2}\right) = 0 \quad \text{e} \quad V(\theta_1) - v_{\theta_1}^*\left(1 - \frac{\alpha}{2}\right) = 0. \quad (7)$$

3 Subamostragem

A ideia subjacente ao método de subamostragem introduzido por Politis e Romano (1994) e Politis et al (1999) é aproximar a distribuição amostral de uma estatística com base no cálculo da mesma em subamostras do conjunto de dados original e não em conjuntos de dados reamostrados como no bootstrap. Assim, sendo \mathcal{P} uma medida de probabilidade, o objectivo é aproximar a distribuição amostral de $(\hat{\theta}_n - \theta(\mathcal{P}))$ pela distribuição empírica dos valores da estatística calculada nos $\binom{n}{b}$ subconjuntos de dimensão $b < n$ de observações amostradas sem reposição. Os valores resultantes do cálculo da estatística em cada uma destas subamostras são convenientemente normalizados de forma a aproximarem a distribuição amostral em causa. Deste modo, a subamostragem constitui um método muito geral para a construção de intervalos de confiança de primeira ordem assintoticamente válidos pois, se b for tal que $b/n \rightarrow 0$ e $b \rightarrow \infty$ quando

$n \rightarrow \infty$, o método é válido sempre que a estatística original, convenientemente normalizada, converge em distribuição para uma determinada distribuição não degenerada.

Note-se por Y_1, Y_2, \dots, Y_{N_n} os $N_n = \binom{n}{b}$ subconjuntos de tamanho b de $Y = \{(t_1, \delta_1), (t_2, \delta_2), \dots, (t_n, \delta_n)\}$ e por $\hat{\theta}_{n,b,i}$ o valor da estatística $\hat{\theta}$ calculada no conjunto de dados Y_i . Seja $J_n(\mathcal{P})$ a distribuição amostral de $\tau_n(\hat{\theta}_n - \theta(\mathcal{P}))$, onde τ_n é uma constante de normalização apropriada. Defina-se também a correspondente função distribuição acumulada como

$$J_n(x, \mathcal{P}) = \text{Prob}_{\mathcal{P}}\{\tau_n(\hat{\theta}_n - \theta(\mathcal{P})) \leq x\} \tag{8}$$

A aproximação subamostragem a $J_n(x, \mathcal{P})$ é então definida por

$$L_{n,b}(x) = \frac{1}{N_n} \sum_{i=1}^{N_n} I\{\tau_b(\hat{\theta}_{n,b,i} - \hat{\theta}_n) \leq x\} \tag{9}$$

onde $I\{E\}$ é a função indicatriz do evento $E = \{\tau_b(\hat{\theta}_{n,b,i} - \hat{\theta}_n) \leq x\}$.

Para cada i , Y_i é uma amostra de dimensão b do verdadeiro modelo \mathcal{P} . Logo, a distribuição exacta de $\tau_b(\hat{\theta}_{n,b,i} - \theta(\mathcal{P}))$ é $J_b(\mathcal{P})$. A distribuição empírica dos N_n valores de $\tau_b(\hat{\theta}_{n,b,i} - \theta(\mathcal{P}))$ constitui uma boa aproximação a $J_n(\mathcal{P})$. Assim, sendo $\theta(\mathcal{P})$ desconhecido, pode-se substituir $\theta(\mathcal{P})$ por $\hat{\theta}_n$ uma vez que $\tau_b(\hat{\theta}_n - \theta(\mathcal{P}))$ é de ordem τ_b/τ_n em probabilidade, assumindo-se que $\tau_b/\tau_n \rightarrow 0$.

Para $\alpha \in [0, 1]$, seja $c_{n,b}(1-\alpha) = \text{inf}\{x : L_{n,b}(x) \geq 1-\alpha\}$. Então, a probabilidade de cobertura assintótica sob \mathcal{P} do intervalo $I_1 = [\hat{\theta}_n - \tau_n^{-1}c_{n,b}(1-\alpha), \infty[$ é $1 - \alpha$.

Um intervalo de confiança bilateral pode ser obtido através da intersecção de dois intervalos unilaterais. Assim, o **intervalo de confiança subamostragem de caudas iguais** de $(1 - \alpha)100\%$ para θ , análogo a I_1 , é

$$I_{ET} = \left[\hat{\theta}_n - \tau_n^{-1}c_{n,b} \left(1 - \frac{\alpha}{2}\right), \hat{\theta}_n + \tau_n^{-1}c_{n,b} \left(\frac{\alpha}{2}\right) \right]. \tag{10}$$

Como aproximação alternativa, também podem ser construídos intervalos de confiança simétricos, sendo, neste caso, a aproximação por subamostragem a $J_{n,|\cdot|}(x, \mathcal{P})$ definida por

$$L_{n,b,|\cdot|}(x) = \frac{1}{N_n} \sum_{i=1}^{N_n} I\{\tau_b|\hat{\theta}_{n,b,i} - \hat{\theta}_n| \leq x\}. \tag{11}$$

Assim, denotando o quantil $(1 - \alpha)$ de $L_{n,b,|\cdot|}$ por $c_{n,b,|\cdot|}(1 - \alpha)$, o **intervalo de confiança subamostragem simétrico** de $(1 - \alpha)100\%$ para θ é dado por

$$I_{sym} = [\hat{\theta}_n - \tau_n^{-1}c_{n,b,|\cdot|}(1 - \alpha), \hat{\theta}_n + \tau_n^{-1}c_{n,b,|\cdot|}(1 - \alpha)]. \tag{12}$$

4 Estudo de simulação

Nesta secção apresentam-se os resultados de um estudo de simulação levado a efeito no *software* estatístico R-Project, que consistiu em aplicar as metodologias bootstrap e subamostragem a dados censurados pelo tempo com o objectivo de calcular intervalos de confiança para o quantil de probabilidade p .

Começaram por gerar-se amostras de dimensão (n) 10, 20 e 30 observações de uma distribuição Weibull com parâmetros de localização $\mu = 0$ e escala $\sigma = 1$. As amostras geradas foram censuradas artificialmente pelo tempo nos valores correspondentes a uma proporção de falhas (p_f) de 0.3, 0.5 e 0.7. Para cada uma destas situações calcularam-se intervalos de confiança a 95% bootstrap e subamostragem para t_p , os quantis de probabilidade (p) 0.1, 0.3, 0.5, 0.7 e 0.9. Na aplicação do bootstrap recorreu-se à metodologia PBSRLLR resultando o intervalo de confiança definido em (7) - $I_{PBSRLLR}$. O valor utilizado para simular a distribuição de $V(t_p)$ definida em (6) foi $B = 999$.

A metodologia subamostragem foi aplicada usando diferentes tamanhos de bloco ($b = 3, 5, 8, 10, 15$ e 20) e calcularam-se intervalos de confiança de caudas iguais I_{ET} , (10), e simétricos I_{sym} , (12). Este procedimento foi repetido 1000 vezes e registou-se o número de intervalos de confiança que, em cada situação, continham o parâmetro de interesse t_p , tendo-se aproximado a probabilidade de cobertura pelo quociente entre o número de intervalos que continham o verdadeiro parâmetro e o número total de intervalos determinados.

Relativamente à metodologia subamostragem, constatou-se que, para a situação em estudo, intervalos de confiança I_{sym} apresentam probabilidades de cobertura superiores aos I_{ET} . Constatou-se ainda que a dimensão de bloco (b) que conduz a melhores aproximações aumenta com a dimensão da amostra a utilizar, resultados estes que estão em conformidade com os estudos de Politis e Romano (1994). Assim, para $n = 10$ obtêm-se melhores aproximações com $b = 3$, para $n = 20$ com $b = 5$ e para $n = 30$ com $b = 10$, pelo que se optou pela apresentação unicamente dos resultados relativos a estas situações.

Nas situações em que a censura é forte, isto é, quando $p_f = 0.3$, verifica-se a existência de um número elevado de conjuntos de dados simulados onde a proporção de falhas é nula, o que inviabiliza o processo de estimação. Assim, optou-se por apresentar apenas as situações em que $p_f \geq 0.5$ e por retirar todos os conjuntos de dados simulados para os quais o número de falhas é inferior a 2, dado que para estas situações as amostras geradas não facilitam o recurso aos métodos numéricos inerentes à estimação de máxima verosimilhança. Mais concretamente, se o número de falhas for 0 não existe estimativa de máxima verosimilhança e, se o número de falhas for 1, em geral, o método iterativo não converge.

As probabilidades de cobertura estimadas dos intervalos de confiança a 95% $I_{PBSRLLR}$ e I_{sym} , para o quantil de probabilidade $p(t_p)$, em função da dimensão da amostra (n) e da proporção de falhas (p_f) apresentam-se na Figura 1.

De um modo geral, as aproximações obtidas, quer através da metodologia bootstrap, quer através da metodologia subamostragem, foram bastante satis-

fatórias, registando-se uma tendência global para o aumento das probabilidades de cobertura quando n aumenta. Ambas as metodologias conduzem a intervalos de confiança que se caracterizam por uma diminuição das probabilidades de cobertura para valores mais elevados do quantil p a estimar.

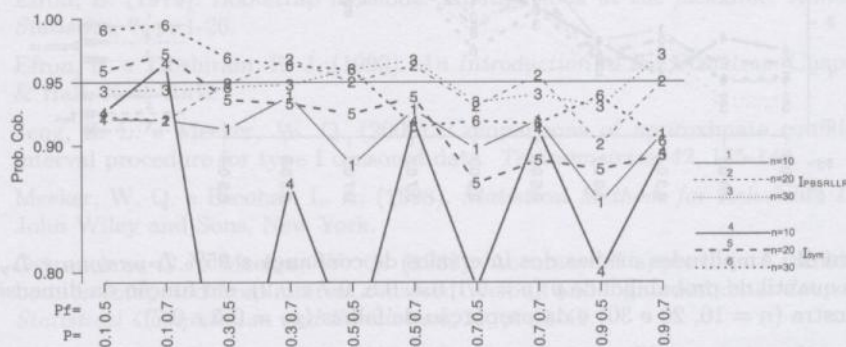


Figura 1: Probabilidade de cobertura estimada para os intervalos de confiança a 95% $I_{PBSRLLR}$ e I_{sym} para o quantil de probabilidade p ($p = 0.1, 0.3, 0.5, 0.7$ e 0.9), em função da dimensão da amostra ($n = 10, 20$ e 30) e da proporção de falhas ($p_f = 0.5$ e 0.7).

São também de assinalar os casos em que p toma valores próximos da proporção de falhas, p_f e o número esperado de falhas é reduzido, os quais traduzem alguma inexactidão na aproximações obtidas. Esta “situação crítica” aparece frequentemente identificada na literatura e, segundo Jeng e Meeker (2000), relaciona-se com o facto de a distribuição do estimador de máxima verosimilhança de t_p depender fortemente de p_f quando $p \approx p_f$.

Note-se que as baixas probabilidades de cobertura dos intervalos de confiança I_{sym} para $n = 10$ podem dever-se ao número disponível de subamostras para a aproximação subamostragem definida em (11), o qual é significativamente mais reduzido que para as restantes situações.

Verifica-se ainda que, se por um lado intervalos de confiança obtidos por subamostragem tendem a apresentar probabilidades de cobertura superiores aos obtidos através do bootstrap, por outro, os intervalos de confiança bootstrap apresentam, em geral, probabilidades de cobertura mais próximas da probabilidade nominal.

Além da probabilidade de cobertura, a amplitude dos intervalos de confiança determinados é também um critério importante para comparar o desempenho das metodologias em estudo. Assim, de dois procedimentos que conduzam a intervalos de confiança com probabilidades de cobertura semelhantes, é melhor aquele que permite obter intervalos de confiança com menores amplitudes.

As amplitudes médias dos intervalos de confiança a 95% $I_{PBSRLLR}$ e I_{sym} , para o quantil de probabilidade p (t_p), em função da dimensão da amostra (n) e da proporção de falhas (p_f) apresentam-se na Figura 2.

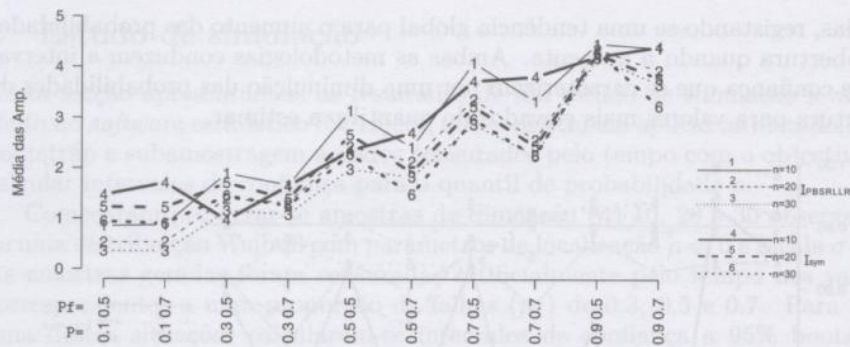


Figura 2: Amplitudes médias dos intervalos de confiança a 95% $IPBSRLLR$ e I_{sym} para o quantil de probabilidade p ($p = 0.1, 0.3, 0.5, 0.7$ e 0.9), em função da dimensão da amostra ($n = 10, 20$ e 30) e da proporção de falhas ($p_f = 0.5$ e 0.7).

Não parece registarem-se diferenças significativas entre as amplitudes médias dos intervalos bootstrap e subamostragem. Destaca-se o facto de se obterem intervalos de confiança com grandes amplitudes para valores elevados do quantil a estimar. O mesmo se passa relativamente à "situação crítica" atrás referenciada, o que conduz, nesses casos, a aproximações muito pobres tanto para a metodologia bootstrap como subamostragem. Regista-se também uma diminuição das amplitudes médias quando n aumenta bem como uma tendência para a diminuição das amplitudes médias quando p_f aumenta.

5 Considerações finais

Neste trabalho procede-se a uma revisão das metodologias bootstrap e subamostragem num contexto de censura Tipo I. Estas técnicas aplicam-se sob condições muito gerais e revelam-se eficazes na determinação de intervalos de confiança para parâmetros populacionais quando não se dispõe de métodos analíticos que permitam inferir acerca do parâmetro estimado.

Num estudo de simulação, aplicam-se as metodologias bootstrap e subamostragem a amostras censuradas artificialmente com o objectivo de se calcularem intervalos de confiança para o quantil de probabilidade p da distribuição do tempo de vida, obtendo-se aproximações bastante satisfatórias para valores moderados do quantil p a estimar. Apesar de o estudo de simulação levado a efeito se basear numa distribuição Weibull, espera-se que as considerações aqui tecidas sejam válidas para outras distribuições da família log-localização-escala. Contudo, seria útil, num trabalho futuro, aplicar as metodologias apresentadas a outros contextos de dados censurados.

Bibliografia

- [1] Davison, A. C. e Hinkley, D. V. (1997). *Bootstrap Methods and their Application*. Cambridge University Press, Cambridge.
- [2] Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics*, 7, p. 1-26.
- [3] Efron, B. e Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. Chapman & Hall, New York.
- [4] Jeng, S. L. e Meeker, W. Q. (2000). Comparisons of approximate confidence interval procedure for type I censored data. *Technometrics*, 42, 135-148.
- [5] Meeker, W. Q. e Escobar, L. A. (1998). *Statistical Methods for Reliability Data*. John Wiley and Sons, New York.
- [6] Ostrouchov, G. e Meeker, W. Q. (1988). Accuracy of approximate confidence bounds computed from interval censored Weibull and lognormal data. *Journal of Statistical Computation and Simulation*, 29, p. 41-49.
- [7] Politis, D. N. e Romano, J. P. (1994). Large sample confidence regions based on subsamples under minimal assumptions. *Annals of Statistics*, 22, p. 2031-2050.
- [8] Politis, D. N., Romano, J. P. e Wolf, M. (1999). *Subsampling*. Springer-Verlag, New York.
- [9] Vander Wiel, S. A. e Meeker, W. Q. (1990). Accuracy and approximate confidence bounds using censored Weibull regression data from accelerated life tests. *IEEE Transactions on Reliability*, R-39, p. 346-351.

Palavras-chave: Interação genética, Modelos estatísticos, Funções de verossimilhança

Abstract: Cordeiro et al. (2011) describe a set of generalized linear models, which have been applied to infer about the nature of the interaction between two genes in binary traits with reduced penetrance: additive, multiplicative, heterogeneity, logistic and probit. As stated by the authors, these models fail to produce truly genetic interpretations. In this work we propose a theoretical framework for reduced penetrance in the case of a single gene, the allelic penetrance approach. Under this approach we develop genetic interaction models, which incorporate the usual concepts of dominance and recessiveness. For illustration purposes, we fit the models through a dataset taken from a genetic mapping of susceptibility to cerebral malaria syndrome in mice.

Keywords: Genetic interaction, Statistical modeling, Allelic penetrance

1 Introdução

No estudo de doenças genéticas, pretende-se descobrir os genótipos de predisposição à doença e a sua relação genética. Uma tarefa esta "facilitada" nos estudos realizados essencialmente pela ação de um único gene, os estudos monogênicos, onde os genótipos conferem o fenótipo. Já quando há um ou vários genes envolvidos de