

## *Aproximações à razão de verosimilhanças*

**Sara Morgado Nunes**

*Escola Superior de Gestão do Instituto Politécnico de Castelo Branco*

**João Renato Sebastião**

*Escola Superior de Gestão do Instituto Politécnico de Castelo Branco*

**Resumo:** Várias aproximações à estatística razão de verosimilhanças têm sido recentemente propostas na literatura com o objectivo de melhorar a inferência sobre um parâmetro populacional  $\theta$ . Neste trabalho consideram-se as aproximações propostas por Barndorff-Nielsen e Severini e apresenta-se uma possível aproximação bootstrap à estatística razão de verosimilhanças. Num estudo de simulação determinam-se intervalos de confiança para o quantil de probabilidade  $p$  através da estatística razão de verosimilhanças e da aproximação de Severini com o objectivo de comparar o desempenho destas metodologias. Na construção dos intervalos de confiança recorre-se à distribuição normal bem como às distribuições bootstrap.

**Palavras-chave:** razão de verosimilhanças, bootstrap, intervalo de confiança, censura à direita.

**Abstract:** Several approximations to the likelihood ratio statistic have been recently proposed in literature in order to improve the inference about a populational parameter  $\theta$ . Here we consider the approximations proposed by Barndorff-Nielsen and Severini and we present a possible bootstrap approximation to the likelihood ratio statistic. In a simulation study we obtain confidence intervals for the  $p$  quantile using the original likelihood ratio statistic and Severini's approximation in order to compare the performance of these methodologies. In the construction of the confidence intervals we use the normal distribution as well as the bootstrap distributions.

**Keywords:** likelihood ratio, bootstrap, confidence interval, right censoring.

## 1 Introdução

A função de verosimilhança foi proposta por Fisher (1922) como sendo uma medida de plausibilidade relativa de vários valores de um dado parâmetro  $\theta$ . O princípio da máxima verosimilhança é hoje amplamente usado na Inferência Estatística, na medida em que constitui um método versátil e muito mais geral que outros conhecidos métodos de estimação. Apesar de as propriedades assintóticas dos métodos baseados na verosimilhança terem sido estudadas durante décadas, continuam a ser uma área em investigação. Na verdade, grande parte do trabalho desenvolvido baseia-se em conjuntos de observações completas, enquanto aspectos relativos à distribuição da estatística razão de verosimilhanças

em contextos de observações incompletas, como acontece, por exemplo, quando os dados em causa são censurados, permanecem menos estudados. Assim, neste trabalho, confere-se especial interesse a dados censurados à direita e, em particular, provenientes de distribuições pertencentes à família de localização e escala.

Considere-se uma amostra aleatória  $Y = \{y_1, y_2, \dots, y_n\}$  de  $n$  observações independentes que pode ser modelada como sendo proveniente de uma distribuição de probabilidade contínua com função densidade de probabilidade  $f(y, \theta)$  onde  $\theta$  é um parâmetro que toma valores em  $\mathbb{R}^d$  e é da forma  $\theta = (\psi, \lambda)$ , sendo  $\psi$  um parâmetro real de interesse e  $\lambda$  um vector de parâmetros perturbadores de dimensão  $d - 1$ . A função de verosimilhança de  $Y$  é definida como sendo uma função de  $\theta$ , proporcional à função densidade do modelo e é dada por

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

Consequentemente, a função log-verosimilhança é dada por

$$l(\theta) = \log L(\theta) = \sum_{i=1}^n \log(f(y_i | \theta)).$$

Maximizando as funções de verosimilhança e log-verosimilhança em ordem a  $\theta$ , obtêm-se estimadores para parâmetros desconhecidos de um dado modelo, a partir de um conjunto de dados disponível. A estimativa de máxima verosimilhança de  $\theta = (\psi, \lambda)$  é notada por  $\hat{\theta} = (\hat{\psi}, \hat{\lambda})$  e obtêm-se resolvendo o sistema de equações  $\frac{\partial l}{\partial \theta} = 0$ .

Na inferência sobre  $\psi$  é frequente recorrer-se à estatística razão de verosimilhanças que é definida como

$$LR(\psi) = \frac{L(\hat{\theta}_\psi)}{L(\hat{\theta})}$$

ou à estatística log-razão de verosimilhanças dada por

$$R^2(\psi) = -2 \log(LR(\psi)) = 2\{l(\hat{\theta}) - l(\hat{\theta}_\psi)\}$$

já que, para grandes amostras, se tem  $R^2(\psi) \sim \chi_1^2$ .

A estatística "signed likelihood" é definida como

$$R(\psi) = \text{sign}(\hat{\psi} - \psi) \sqrt{2\{l(\hat{\theta}) - l(\hat{\theta}_\psi)\}}, \quad (1)$$

onde  $\hat{\theta}_\psi$  denota o estimador de máxima verosimilhança de  $\theta$  para  $\psi$  fixo e  $\text{sign}(x) = |x|/x$ .

A estatística  $R(\psi)$  é assintoticamente normal standard com erro de ordem  $O(n^{-1/2})$ . Assim, os limites de um intervalo de confiança a  $100(1 - \alpha)\%$  para  $\psi$  com base em  $R(\psi)$  são dados pela solução das equações

$$R(\psi) - z \left( \frac{\alpha}{2} \right) = 0 \quad \text{e} \quad R(\psi) + z \left( \frac{\alpha}{2} \right) = 0. \quad (2)$$

Porém, esta aproximação pode ser pobre na presença de parâmetros perturbadores ou se a amostra em causa for de pequena dimensão dado que a estatística "signed-likelihood" não se comporta como uma verdadeira verosimilhança e logo não possui todas as propriedades desta função. Assim, com o objectivo de melhorar a inferência sobre  $\psi$ , durante as duas últimas décadas, têm sido propostos na literatura diversos procedimentos baseados na verosimilhança que apresentam grande precisão, mesmo em amostras de dimensão reduzida, entre os quais se destacam os ajustamentos à estatística  $R(\psi)$  propostos por Barndorff-Nielsen (1986,1991) e Severini (1999).

## 2 Aproximações à razão de verosimilhanças

Com o objectivo de melhorar a convergência da distribuição da estatística "signed-likelihood" para a distribuição normal, Barndorff-Nielsen (1986,1991) e Severini (1999) propõem modificações da forma

$$\tilde{R}(\psi) = R(\psi) + \log \left( \frac{U(\psi)}{R(\psi)} \right) \frac{1}{R(\psi)}, \quad (3)$$

onde  $R(\psi)$  é a estatística "signed-likelihood" definida em (1) e  $U(\psi)$  é um valor real que pode ser obtido de diferentes formas. O factor de ajustamento  $U(\psi)$  de Barndorff-Nielsen (1986,1991) é definido como

$$U_{BN}(\psi) = \frac{\begin{vmatrix} l_{;\hat{\psi}}(\hat{\theta}) - l_{;\hat{\psi}}(\hat{\theta}_\psi) & l_{\lambda;\hat{\psi}}(\hat{\theta}_\psi) \\ l_{;\hat{\lambda}}(\hat{\theta}) - l_{;\hat{\lambda}}(\hat{\theta}_\psi) & l_{\lambda;\hat{\lambda}}(\hat{\theta}_\psi) \end{vmatrix}}{\sqrt{|j_{\theta\theta}(\hat{\theta})||j_{\lambda\lambda}(\hat{\theta}_\psi)|}}. \quad (4)$$

Note-se que

$$l_{;\hat{\psi}}(\theta) = \frac{\partial}{\partial \hat{\psi}} l(\theta), \quad l_{;\hat{\lambda}}(\theta) = \frac{\partial}{\partial \hat{\lambda}} l(\theta), \quad l_{\lambda;\hat{\psi}}(\theta) = \frac{\partial}{\partial \lambda} l_{;\hat{\psi}}(\theta), \quad l_{\lambda;\hat{\lambda}}(\theta) = \frac{\partial}{\partial \lambda} l_{;\hat{\lambda}}(\theta)$$

$$j_{\theta\theta}(\theta) = \begin{bmatrix} j_{\psi\psi}(\theta) & j_{\lambda\psi}(\theta) \\ j_{\psi\lambda}(\theta) & j_{\lambda\lambda}(\theta) \end{bmatrix} = \begin{bmatrix} \frac{\partial^2}{\partial \psi^2} l(\theta) & \frac{\partial^2}{\partial \psi \partial \lambda} l(\theta) \\ \frac{\partial^2}{\partial \psi \partial \lambda} l(\theta) & \frac{\partial^2}{\partial \lambda^2} l(\theta) \end{bmatrix},$$

sendo  $\lambda$  o vector de parâmetros perturbadores e  $\hat{\lambda}$  o respectivo estimador de máxima verosimilhança.

De facto, verifica-se que a distribuição de  $\tilde{R}(\psi)$  com  $U_{BN}(\psi)$  é assintoticamente normal standard com erro de ordem  $O(n^{-3/2})$ , o que conduz a testes de significância e limites de confiança bastante precisos (Barndorff-Nielsen, 1991).

Porém, o cálculo do factor de ajustamento  $U_{BN}(\psi)$  pode tornar-se difícil uma vez que envolve a determinação de um conjunto de derivadas em que é necessário recorrer a uma estatística ancilar  $a$  para factorizar a função log-verosimilhança na forma  $l(\theta) = l(\theta; \hat{\theta}|a)$ , sendo na maior parte dos casos, limitado aos modelos da família exponencial. Com o objectivo de desenvolver uma aproximação que seja facilmente calculada para um leque mais vasto de modelos, Severini (1999) propõe que o cálculo do factor de ajustamento  $U(\psi)$  se baseie apenas na função log-verosimilhança e nas respectivas derivadas parciais em relação aos parâmetros do modelo e não em relação aos estimadores de máxima verosimilhança, como acontecia com  $U_{BN}(\psi)$ .

Assim, o factor de ajustamento proposto por Severini (1999) é dado por

$$U_S(\psi) = \frac{\begin{vmatrix} \hat{\Delta}_\psi(\psi) & \hat{\Sigma}_{\lambda\psi}(\psi) \\ \hat{\Delta}_\lambda(\psi) & \hat{\Sigma}_\lambda(\psi) \end{vmatrix}}{\sqrt{|j_{\theta\theta}(\hat{\theta})||j_{\lambda\lambda}(\hat{\theta}_\psi)|}}, \quad (5)$$

onde as matrizes  $\hat{\Delta}_\psi(\psi)$  e  $\hat{\Delta}_\lambda(\psi)$  são as componentes de  $\hat{\Delta}(\psi)$ , definida por

$$\hat{\Delta}(\psi) = j(\hat{\theta})i(\hat{\theta})^{-1} \left[ \hat{Q}(\hat{\theta}_\psi; \hat{\theta}) - \hat{Q}(\hat{\theta}; \hat{\theta}) \right],$$

e as matrizes  $\hat{\Sigma}_{\lambda\psi}(\psi)$  e  $\hat{\Sigma}_\lambda(\psi)$  são as componentes de  $\hat{\Sigma}(\psi)$ , definida por

$$\hat{\Sigma}(\psi) = j(\hat{\theta})i(\hat{\theta})^{-1} \hat{I}(\hat{\theta}_\psi; \hat{\theta}),$$

com

$$i(\hat{\theta}) = \hat{I}(\hat{\theta}; \hat{\theta}), \quad \hat{Q}(\theta; \theta_0) = \sum_{i=1}^n l_\theta^{(j)}(\theta_0) l^{(j)}(\theta) \quad \text{e} \quad \hat{I}(\theta; \theta_0) = \sum_{i=1}^n l_\theta^{(j)}(\theta_0) l_\theta^{(j)}(\theta)^T,$$

sendo  $l^{(j)}(\theta)$  a função de log-verosimilhança relativamente à  $j$ -ésima observação e  $l_\theta^{(j)}(\theta)$  a matriz das derivadas parciais de  $l^{(j)}(\theta)$ .

Os limites de um intervalo de confiança a  $100(1 - \alpha)\%$  para  $\psi$  com base em  $\tilde{R}(\psi)$  são dados pela solução das equações

$$\tilde{R}(\psi) - z \left( \frac{\alpha}{2} \right) = 0 \quad \text{e} \quad \tilde{R}(\psi) + z \left( \frac{\alpha}{2} \right) = 0, \quad (6)$$

sendo  $\tilde{R}(\psi)$  obtido a partir do factor de ajustamento  $U_{BN}(\psi)$  ou  $U_S(\psi)$  definidos em (4) e (5), respectivamente.

Neste trabalho propõem-se duas aproximações bootstrap à estatística razão de verosimilhanças obtidas com base nas distribuições empíricas das funções  $R(\psi)$  e  $\tilde{R}(\psi)$  definidas em (1) e (3), respectivamente.

Seja  $F(\psi)$  uma função que pode representar  $R(\psi)$  ou  $\tilde{R}(\psi)$ . A distribuição empírica de  $F(\psi)$  é calculada recorrendo a  $F_1^*, F_2^*, \dots, F_B^*$ , valores calculados da seguinte forma:

- para  $i = 1, \dots, B$  seleccionam-se aleatoriamente com reposição  $n$  valores da amostra original  $Y$ , formando-se assim uma nova amostra  $Y_i^*$ ;
- recorrendo à amostra  $Y_i^*$  calcula-se o valor de  $F_i^* = F(\hat{\psi})$ , sendo  $\hat{\psi}$  o estimador de máxima verosimilhança da amostra original.

O *quantil*  $-\alpha$ ,  $v^*(\alpha)$ , da distribuição empírica de  $F(\psi)$  é definido como sendo o  $(B+1)\alpha$ -ésimo valor ordenado de  $F_i^*$ .

Assim, os limites de um intervalo de confiança bootstrap a  $100(1-\alpha)\%$  para  $\psi$  são dados pela solução das equações

$$F(\psi) - v^* \left(1 - \frac{\alpha}{2}\right) = 0 \quad \text{e} \quad F(\psi) - v^* \left(\frac{\alpha}{2}\right) = 0. \quad (7)$$

### 3 Estudo de simulação

A maior dificuldade em generalizar os procedimentos de estimação usuais para dados censurados deve-se ao facto de, sob censura, a função log-verosimilhança e as suas derivadas serem uma mistura de variáveis aleatórias com uma parte discreta e uma parte contínua, tornando o seu cálculo muito complexo. Assim, num estudo de simulação levado a efeito no *software* estatístico R-Project estudou-se a fiabilidade dos intervalos de confiança definidos por Severini (1999) em diferentes situações. Para tal, recorreu-se à estatística "signed-likelihood" definida em (1), à aproximação definida em (3) com o factor de ajustamento proposto por Severini (1999) conforme se definiu em (5) e às respectivas aproximações bootstrap para a determinação de intervalos de confiança para o quantil de probabilidade  $p$  ( $t_p$ ) num contexto de censura à direita, onde os dados são provenientes de uma distribuição Weibull, cuja função densidade de probabilidade se define por

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

sendo  $\phi(z) = \exp(z - \exp(z))$  e  $t_p$  definido por

$$t_p = \exp(\mu + \Phi^{-1}(p)\sigma),$$

onde  $\Phi^{-1}(p) = \log(-\log(1-p))$ .

Na situação concreta de a distribuição em causa ser Weibull, a obtenção do factor de ajustamento  $U(\psi)$  em (3) fica inviabilizada devido à inexistência de uma estatística ancilar para a factorização da função log-verosimilhança  $l(\theta) = l(\theta; \hat{\theta}|a)$ , revelando-se assim bastante útil a aproximação proposita por Severini em (5).

Geraram-se amostras de dimensão  $n=10$  e  $30$  de uma distribuição Weibull com parâmetros de localização  $\mu = 0$  e escala  $\sigma = 1$ , censuradas artificialmente pelo tempo nos valores correspondentes a uma proporção de falhas  $p_f=0.1, 0.3, 0.5, 0.7$  e  $0.9$ . O mecanismo de obtenção das amostras censuradas foi o seguinte:

1. Considerar  $r = n.p_f$  e  $U_{(0)} = 0$
2. Gerar aleatoriamente  $r$  observações  $U_1, \dots, U_r$  de uma variável com distribuição Uniforme(0,1)
3. Calcular as correspondentes estatísticas ordinais

$$U_{(i)} = 1 - (1 - U_{(i-1)})(1 - U_i)^{\frac{1}{n-i+1}}, \quad i = 1, \dots, r$$

4. A amostra aleatória é definida por

$$X_{(i)} = \Phi^{-1}(U_{(i)}), \quad i = 1, \dots, r$$

Para cada das amostras geradas calcularam-se intervalos de confiança a 95% definidos em (2), (6) e (7) para os quantis de probabilidade  $p=0.1, 0.3, 0.5, 0.7$  e  $0.9$ . O intervalo de confiança com base na estatística "signed-likelihood" original definido em (2) é notado por  $I_R$ , o intervalo obtido com base na aproximação  $\tilde{R}(\psi)$  e definido em (6) é notado por  $I_{R_S}$  e os intervalos bootstrap com base em  $R(\psi)$  e  $\tilde{R}(\psi)$  definidos por (7) por  $I_{R^*}$  e  $I_{R_S^*}$ , respectivamente. A estatística  $\tilde{R}(\psi)$  foi sempre calculada a partir do factor de ajustamento  $U_S(\psi)$  proposto por Severini (1999) e definido em (5). Na obtenção da distribuição empírica de  $R(\psi)$  e  $\tilde{R}(\psi)$  usaram-se  $B = 999$  réplicas bootstrap. 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.

As probabilidades de cobertura estimadas dos intervalos de confiança a 95%  $I_R, I_{R_S}, I_{R^*}$  e  $I_{R_S^*}$  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.

Os resultados obtidos foram bastante satisfatórios mesmo para  $n = 10$ , registando-se uma tendência global para o aumento das probabilidades de cobertura quando  $n$  aumenta. Os intervalos  $I_{R_S}$  e  $I_{R_S^*}$  tendem a apresentar probabilidades de cobertura mais próximas da probabilidade de cobertura nominal que os intervalos  $I_R$  e  $I_{R^*}$ . Para  $p_f = 0.3$  obtêm-se aproximações muito pobres o que se deve ao facto de a censura ser forte dado que se verifica a existência

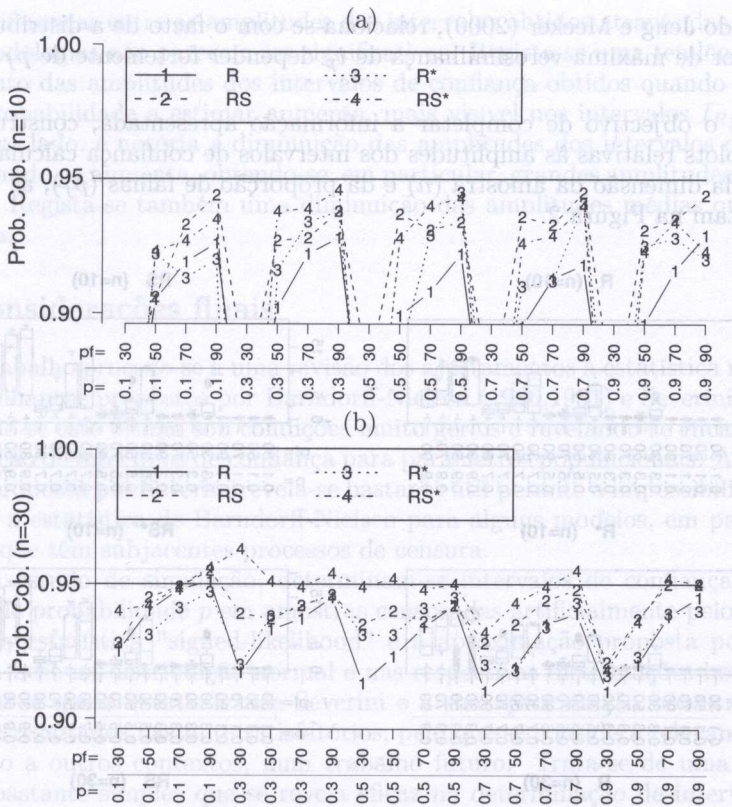


Figura 1: Probabilidade de cobertura estimada para os intervalos de confiança a 95%  $I_R$ ,  $I_{RS}$ ,  $I_{R^*}$  e  $I_{RS^*}$  para o quantil de probabilidade  $p$  ( $p = 0.1, 0.3, 0.7$  e  $0.9$ ), em função da proporção de falhas ( $p_f = 0.3, 0.5, 0.7$  e  $0.9$ ) para  $n = 10$  (a) e  $n = 30$  (b).

de um número elevado de conjuntos de dados simulados onde a proporção de falhas é nula, o que inviabiliza o processo de estimação. Para atenuar esta situação, optou-se 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. 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 nas 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  $p \approx p_f$ .

Com o objectivo de completar a informação apresentada, construíram-se as box-plots relativas às amplitudes dos intervalos de confiança calculados, em função da dimensão da amostra ( $n$ ) e da proporção de falhas ( $p_f$ ), as quais se apresentam na Figura 2.

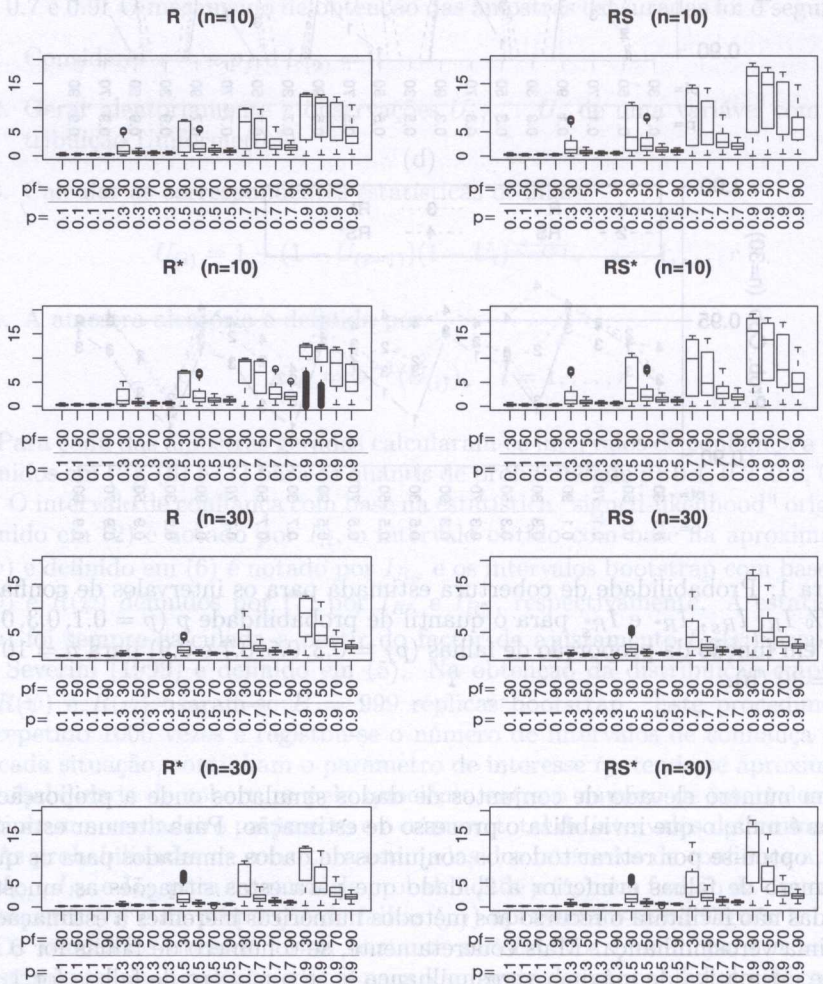


Figura 2: Box-plots das amplitudes dos intervalos de confiança a 95%  $I_R$ ,  $I_{RS}$ ,  $I_{R^*}$  e  $I_{RS^*}$  para o quantil de probabilidade  $p$  ( $p = 0.1, 0.3, 0.7$  e  $0.9$ ), em função da proporção de falhas ( $p_f = 0.3, 0.5, 0.7$  e  $0.9$ ) para  $n = 10$  e  $n = 30$ .

As diferenças entre as amplitudes dos intervalos obtidos através das diferentes metodologias não parecem ser significativas. Regista-se uma tendência para o aumento das amplitudes dos intervalos de confiança obtidos quando o quantil de probabilidade a estimar aumenta, mais visível nos intervalos  $I_{R_S}$  e  $I_{R_S^*}$ . Por outro lado, é notória a diminuição das amplitudes dos intervalos de confiança quando  $p_f$  aumenta, obtendo-se, em particular, grandes amplitudes quando  $p_f \leq p$ . Regista-se também uma diminuição das amplitudes médias quando  $n$  aumenta.

#### 4 Considerações finais

Neste trabalho procede-se a uma revisão dos ajustamentos à estatística razão de verosimilhanças propostos por Barndorff-Nielsen (1986,1991) e Severini (1999), aplicando-se este último sob condições muito gerais e revelando-se eficaz na determinação de intervalos de confiança para parâmetros populacionais. A aproximação proposta por Severini revela-se bastante útil perante a impossibilidade de calcular a estatística de Barndorff-Nielsen para alguns modelos, em particular aqueles que têm subjacentes processos de censura.

Num estudo de simulação, determinam-se intervalos de confiança para o quantil de probabilidade  $p$  em amostras censuradas artificialmente pelo tempo, usando a estatística "signed-likelihood" e a aproximação proposta por Severini com base na distribuição normal e nas respectivas distribuições bootstrap. Constata-se que a estatística de Severini e a sua aproximação bootstrap conduzem a resultados bastante satisfatórios, pelo que se considera relevante a sua aplicação a outros contextos, num trabalho futuro. Trata-se de uma aproximação bastante simples que se revela eficaz na determinação de intervalos de confiança para qualquer estatística, ainda que para amostras de pequena dimensão.

#### Referências

- [1] Barndorff-Nielsen, O. E. (1986). Inference on full and partial parameters, based on the standardized signed log likelihood ratio. *Biometrika*, Vol. 73, p. 307-322.
- [2] Barndorff-Nielsen, O. E. (1991). Modified signed log likelihood ratio. *Biometrika*, Vol. 78, p. 557-563.
- [3] Fisher, R. A. (1922). On mathematical foundation of theoretical statistics. *Philos. Trans. Roy. Soc. Series A*, Vol. 222, p. 309-368.
- [4] Jeng, S. L. e Meeker, W. Q. (2000). Comparisons of approximate confidence interval procedure for type I censored data. *Technometrics*, Vol. 42, p. 135-148.
- [5] Severini, T. A. (1999). An empirical adjustment to the likelihood ratio statistic. *Biometrika*, Vol. 86, p. 235-247.